Mesoscopic Physics

an introduction

TN385 (Form. C29c)

C. Harmans

Version 2, November 1997

TU Delft

20.-

Table of contents

1. Introduction

2. Diffusive transport: classical and quantum regime

3. Classical ballistic transport

4. Quantum ballistic transport

5. Aharonov Bohm effect and persistent currents

6. Transport in normal-superconductor systems

7. Coulomb interaction and charging effects

8. Optical properties of confined systems *(in preparation)*

9. Thermodynamic properties of clusters

10. Tunneling and scanning probe methods

Appendices

A. Nanostructure fabrication techniques (*in preparation*)

B. Measuring methods in mesoscopic physics *(in preparation)*

General information

This text is the second version of the "collegediktaat" which accompanies the two-semester course on Mesoscopic Physics offered by the Department of Applied Physics of the Delft University of Technology. The course, coded TN 385 (formerly C29c), is configured to be an introduction into an area of solid state physics which has seen a very rapid development during the last decade, a development which is still continuing and forming new branches. This text contains the material that was felt to be most appropriate for understanding and appreciating the developments in the field. It intends to serve as a survey of the field, allowing advanced undergraduates and starting graduates to obtain a global overview.

The course requires an introductory level knowledge of quantum mechanics, thermodynamics and solid state physics. Some knowledge of superconductivity is beneficial but not essential. The main goal of the course is to introduce the physical concepts underlying the phenomena in this field. This is persuaded by the use of transparent physical models to describe the various phenomena. These models are extensively complemented by recent experimental results, highlighting the main aspects introduced in the model but at the same time stressing the limitations by identifying deviations found experimentally. While thoroughly discussing the formalisms essential for understanding the subject it avoids lengthy and highly technical theoretical derivations. Where possible suggestions for further reading are given providing convenient paths for an in-depth study.

In preparing this second version the author has greatly benefitted from input received from (undergraduate) students, A/OIO's (graduate students) as well as staff members. He still would be highly indebted if any further suggestions for improvement would be made known to him.

Delft, November 1997

1. Introduction to the field of Mesoscopic Physics

Introduction

The physics of solid state materials or condensed matter is of central importance within science as a whole. Being "born" during the thirties of the 20th century, one may assume it to be fully mature by now, without significantly new discoveries being made any more. In contrast, condensed matter physics shows all signatures of genuine prosperity, including a marked growth. This growth is-driven by inputs from fundamental as well as applied oriented origin.
Condensed matter physics concentrates on phenomena occurring in solids, studying thermal, mechanical, optical, electrical, magnetic, . properties. In general it assumes the system of interest to be large, i.e. to contain a large number of "unit building blocks", mostly atoms. This case, indicated as the thermodynamic limit, can be addressed by either assuming the building blocks to be arranged in perfect order as to form regular lattices, or to be organised in a complete random manner. In these two limits the system can be modelled in such a way that mathematical techniques are available to evaluate the physical quantities of interest. In this way only the properties of bulk material is obtained. Note that a bulk system not necessarily implies that only systems that extend into three dimensions are allowed: surfaces or interfaces are also included (forming 2-dimensional bulk systems) provided these are extended over many "units cells" in two dimensions.

Mesoscopic physics forms the bridge between the macroscopic world of bulk materials and the microscopic world of atoms and molecules, the behaviour of which usually requires more involved methods to reveal their secrets. Mesoscopic physics is interested in questions such as: how will electrons start to behave if only very few of them are available; from what size of a piece of material the wave-nature, governed by quantum mechanics, becomes "averaged away" thus recovering the more common classical behaviour; under what conditions will the "granularity" of the charge become apparent; and many more.
The text nearly exclusively concentrates on phenomena of an electrical nature, i.e. driven from the free electrons in the system.

In this first chapter an introduction to the field is given. The first section (1.1) provides a few examples to position the field within physics in general and in solid state physics in particular. It provides a first opportunity to "taste the flavour" of the field. Most of these examples will return in forthcoming chapters where they will be discussed in greater depth. Next, section 1.2 introduces a number of basic properties of conducting systems, in particular concentrating on various length scales and energy scales that govern the transport of free electrons in a solid. In addition the role of the dimensionality of such systems is introduced and explained. Section 1.3 finally defines the various regimes for electron transport, and how a solid material can be manipulated in such a way that the required type of transport (or transport regime) can be realised.

1.1 A global positioning: some simple examples

In this section a number of examples of mesoscopic systems will be presented. These examples are chosen such that the key aspects of the phenomenon can be explained in a simple and intuitive manner. In this way these cases act as a global introduction to the type of phenomena that are of interest in this field. Most of the examples will be reinvestigated in one of the following chapters, where these will become discussed in considerably more detail and depth.

1.1.a. *Electron bound states in gold clusters*

Our first example concerns so called *clusters of atoms*. In single, isolated atoms and molecules electrons are bound to the nuclei in well-defined and stable orbits. These stationary orbits are an immediate consequence of the quantum rules operating at this level of size. Typically the size of these objects range from some 50 pm (1 pm $=10^{-12}$ m) for the smallest atoms to 0.3 nm for the larger ones and up to 10 nm (1 nm $=10^{-9}$ m) for large organic molecules. The stationary nature of the electron orbits implies that a bound electron resides in a well defined *electronic state* with a well determined energy associated with it. Such an eigenstates is completely characterised by a unique set of quantum numbers. In a system containing more than one electron, the electrons fill up the available eigenstates of the system in increasing energy, with one electron being allowed for each state. This "one electron per state" is an immediate consequence of the well known Pauli exclusion principle, a basic rule of play for all half-spin particles (thus including electrons) in nature. So, in a system like an atom containing a given number of electrons which is in the ground state, there is an energy beyond which the available states are unoccupied and contain no electrons.
Atoms and molecules can be forced out of the ground state and become excited. In that case one electron in a certain state/orbit increases its energy by going to an orbit associated with a state of larger energy; this process can be driven by the absorption of a quantum of electron magnetic radiation, i.e. a photon. Evidently absorption only will occur if the energy of the photon matches the energy difference between the initial (or equilibrium) and the final (or excited) state. For the case of electrons attached to an atom the energy level differences for the high-lying states (i.e. the outermost electrons) have values of typically 1 eV (1 eV= $1.602*10^{-19}$ J).
In a piece of solid material individual atoms are arranged (ideally) such as to form a regular lattice. This happens by the atoms forming mutual bonds, actually by sharing some of their outermost electrons. From solid state physics we know that, if the material formed is a metal, then these outermost electron(s) leaves "its" core (= nucleus + the remaining bound electrons) all the way and is allowed to move freely throughout the whole piece of material. These free electrons actually make the material to be metallic, showing electrical conduction.
Now we pose the question as to what will happen if we shrink the piece more and more. At some moment the lump of material (or grain) may start to look similar to a big molecule. This means that the free electrons feel the boundaries of the grain and become ordered in well determined orbits, forming eigenstates with associated eigenenergies and a set of quantum numbers. Note that the stationary orbits of these states extend

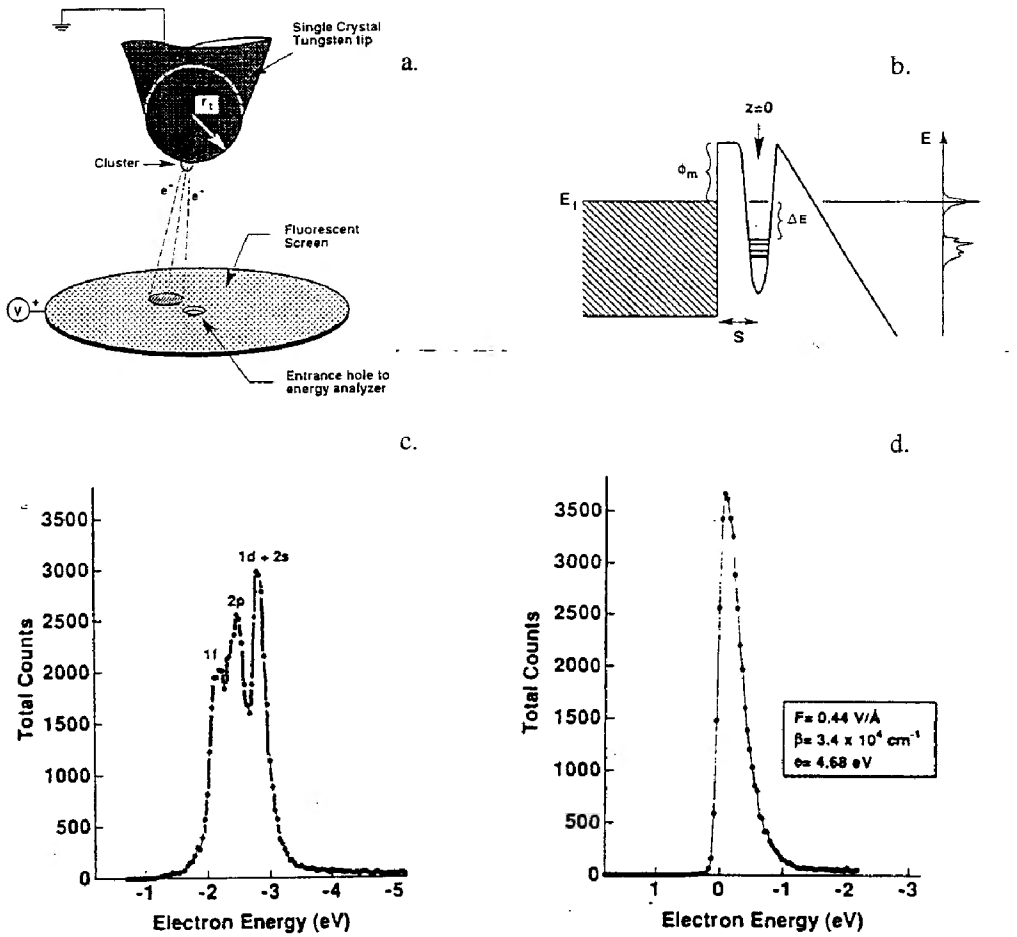


Figure 1.1: Discrete electrons states in a ~ 1 nm gold cluster; a. diagram of the set-up for field emission; b. energy diagram of the cluster at the tip; c. energy distribution of emitted electrons from the cluster; d. distribution from the tip

throughout the whole grain! The precise shape of an individual orbit depends on the shape of the grain, and consequently the same holds for the eigenenergy of the state thus formed. This is an important general notion: *the nature of the eigenstates of a system is determined by its shape*. We will return to this in more detail in chapter 9.
Grains of this size are known in the literature as (quantum) *clusters*.
Recently this phenomena has indeed been found experimentally (Lin et.al., Phys. Rev. Lett. 67, 477 (1991))*. Figure 1.1 shows the most relevant pictures related to this experiment on a gold cluster of ~1 nm diameter.

* Wherever a reference is given to some paper it is worthwhile to look into it in the library!

Question: estimate the number of Au atoms contained in this cluster; how many *free* electrons will there be in the cluster?
In fig 1.1a the set-up is shown schematically. The cluster, attached to a *very sharp needle*, is held at zero potential. A nearby screen, with a small hole in its centre, is put at a large positive voltage. The sharpness (or more precisely the curvature) of the tip leads to a very large electric field in the vicinity of the tip, resulting in *field-emission* of electrons from the tip. If a small object is positioned at the end of the tip, the field will be largest there and so electrons will be field-emitted from this object. This is shown in fig 1.1a. In fig 1.1b an energy diagram for the tip is displayed. The cluster is modelled as a potential well at position z=0, with the bound electron states shown inside the well. Electrons can be extracted from these bound states and these may leave the cluster via tunnelling through the triangular barrier at the right. Quantum mechanics tells us that the probability for this tunnelling depends on the height of the barrier (and so the binding of the electron) and its width. This width is controlled by the slope of the right side of the barrier, and so by the strength of the electric field (i.e. the accelerator voltage between screen and needle).
If an electron tunnels out its *total energy* is given by the sum of the accelerator voltage *and the bound state eigenenergy*; note that this total energy will be *less* than the accelerator voltage! The emitted electrons are accelerated towards the screen. By manipulating the position of the hole relative to the tip the high-energy electrons may fly through, entering an energy analyser located behind the screen. In this analyser each electron is counted and its energy is determined, leading to a spectrum of (number-of-electrons) <=> energy. As the electrons from the *different* bound states have *different* total energies they will show up in the spectrum as different peaks. This is exactly what is shown in fig. 1.1c and d.
In particular d. shows what is found for a clean tip, where no particular structure in the spectrum is seen. In contrast fig. 1.1c shows the spectral distribution for the Au cluster: a clear split-peak-shaped distribution (at a *lower total* energy, as anticipated) is evident. This result nicely demonstrates the bound state eigenenergies of a metallic cluster.

1.1.b. *Shell structure and magic numbers in clusters*

The second example to be discussed, somehow complements the first one and concerns the occurrence of a *shell structure* and *associated magic numbers in metallic clusters*. This can be understood rather simply starting from purely classical arguments.
If we build clusters from identical particles we readily find that at certain numbers of particles the cluster will be highly regular and symmetric. Fig 1.2 shows this in a nice way. In such cases all positions available within one particular "shell" are filled; the particle numbers associated with these cases are denoted as *magic numbers*.
Now we switch to quantum mechanics. Basically also for the quantum analogue of such high-symmetry cases the system is said to have completely filled shells, characterised by the principle quantum number n. From quantum mechanics we know that adding a particle to the next higher shell usually takes more energy than adding it to the same shell. This implies that clusters with only filled shells (and thus with a total number of particels given by a magic number) will have a lower total energy as compared to non-magic counterparts. This makes is not unlikely that, if clusters are allowed to condense from (e.g.) a vapour of free atoms, that there will be a preference to form clusters with filled

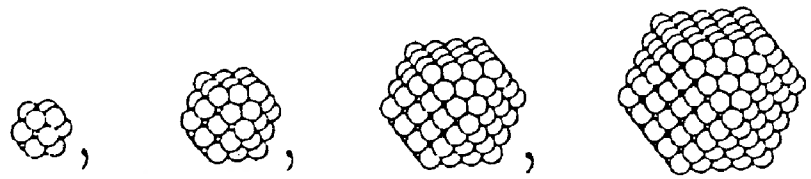


Figure 1.2: Sequence of magic number clusters

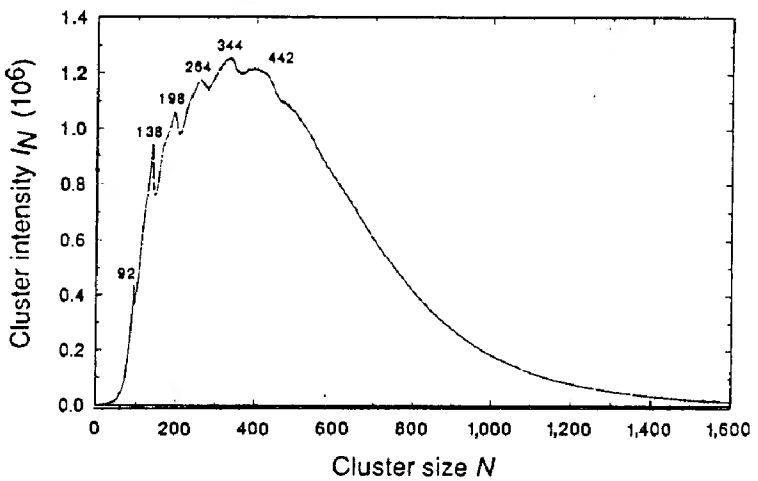


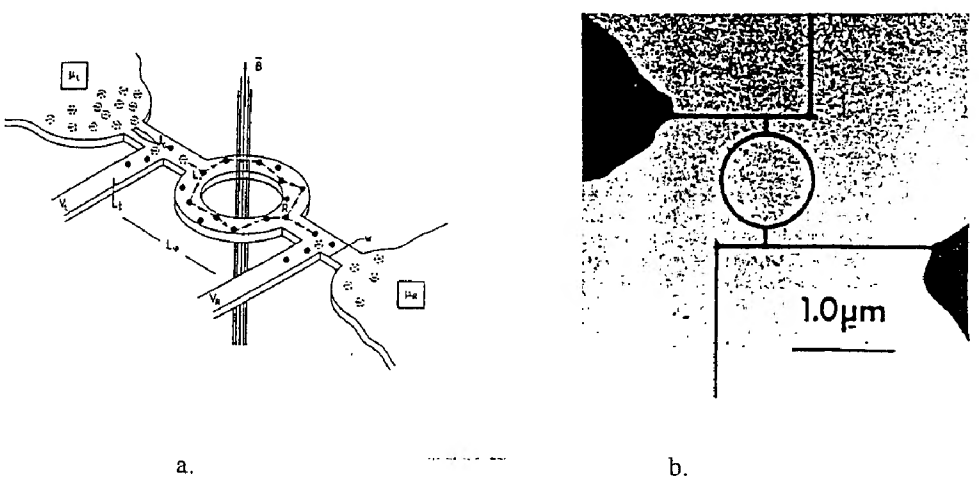
Figure 1.3: Mass spectrum versus cluster size

shells. This would lead to a distribution of the masses of the clusters with preference for the magic numbers. This is what is found experimentally.
Figure 1.3 shows the results of an experiment where clusters are allowed to form in free space, after which their masses are measured in a mass-spectrometer (Nature 333, 734 (1991)). Clearly the spectrum shows peaks of increased intensity, demonstrating the preference for certain masses.

1.1.c. *Electron interference in solid state rings*

The next example concerns the *transport of electrons* in a *small structure*. In particular we will see that in a ring-shaped conductor (with two leads attached) the electrons may show wave-interference phenomena, which can be controlled by an external magnetic field. This is visible in the resistance (or conductance) of the ring+leads.
Interference effects require wave properties, and thus we have to define which wavelength will be relevant for free electrons in a solid. From introductory solid state physics we know that in an ideal crystalline system the translation invariance of the lattice leads to the formation of energy bands, with a finite number of states available per band

(more precisely: a finite density of states per band). In case of a metal these bands are only partially filled by electrons, up to a *maximum* energy, denoted as the Fermi energy. Electrical conduction can take place because of this partial filling of the topmost band. Associated with the Fermi energy there is a *maximum* wavevector k_F and so a *minimum* wavelength $\lambda_F=2\pi/k_F$, the Fermi wavelength. As in a conductor only electrons with energies close to the Fermi energy contribute to the conduction also only the Fermi wavelength will be of importance for the discussion that follows.
In the majority of macroscopic cases the fact that electrons do have a wavelength (and so an associated phase) is not immediately evident. Figure 1.4a however shows a configuration where this effect may become apparent. The sample is ring shaped, allowing the current to take both arms in traversing it from the top left to the lower right large contact); the added other two leads need not concern us now (see Appendix B for some details on 2- and 4-terminal configurations). Now we have to realise that electrons behave as waves. This implies that an electron(-wave) entering the ring at one side will *split and the resulting (partial) waves will travel the two arms of the ring* and *reunite* at the other end. The way this reunion occurs is defined by adding the two partial wave amplitudes *vectorially* and not as scalars. This implies that not only the partial amplitudes will determine the amplitude of the sum wave, but just as much the (relative) *phase* of the two components. This phase-dependent summation effectively leads to *electron interference.*
From electrodynamics and quantum mechanics we know that the evolution of the phase of a charged particle (like an electron) depends on its velocity but, in addition, it is



a. b.

Figure 1.4.: metallic ring structure for the Aharonov-Bohm experiments a.: schematic diagram of the ring; b.: SEM photograph of Au ring of approximately 800 nm diameter, evaporated onto a non-conducting substrate.

affected by a magnetic field B. More specifically the particle travelling along a path l acquires an additional phase shift which can be obtained from the *vectorpotential* A (associated with the field as $B = \nabla \wedge A$) given by:

$$\Delta\phi = \frac{e}{\hbar}\int A.dl \qquad (1.1)$$

Note that $A.dl$ represents a *vectorproduct*. Getting back to our fig 1.4a it is now evident that electron waves travelling the two different branches of the ring will also acquire *different* additional phases due to the magnetic field via the different values of $A.dl$ for the two half-rings. As will be derived in chapter 5 this phase difference is simply determined by the total magnetic flux Φ enclosed by the area S of the ring:

$$\Delta\phi = \frac{e}{\hbar}BS = 2\pi\frac{\Phi}{\Phi_0} \qquad (1.2)$$

Φ_0 is called the *flux quantum*, given by h/e. The additional phase $\Delta\phi$ commonly is called the *Aharonov-Bohm (AB) phase*, after the two persons who first realised that effects due to this phase should be experimentally accessible.
Now what will happen to the phase if we vary the magnetic field? Assume for the sake of simplicity that at B=0 the electron waves interfere *constructively*, i.e. the two vectors arriving at the point of reunion have to be added "in line" and so will lead to a *maximum* in the total sum wave. Stated differently, this implies a *maximum probability* for the electron wave to travel through the ring or equivalently a *minimum in the resistance* of the ring. Now let us increase the field such as to add half a flux quantum to it. Eq. (1.2) then indicates that the AB phase changes by π, and so the former constructive interference will change to a *destructive* one. Evidently this implies that the two vectors are in anti-parallel and so will yield a small sum. So, the the resistance will now be at a maximum. Increasing the field further to get one flux quantum in the ring will lead to an AB phase of 2π, a value which is evidently indistinguishable from a phase difference of zero: so again a resistance minimum. Summarising the whole sequence shows that the resistance *oscillates* between a maximum and minimum value, due to the interference of the electron waves. The period of this oscillation is given by $\Delta B = \Phi_0/S$, i.e. one cycle per

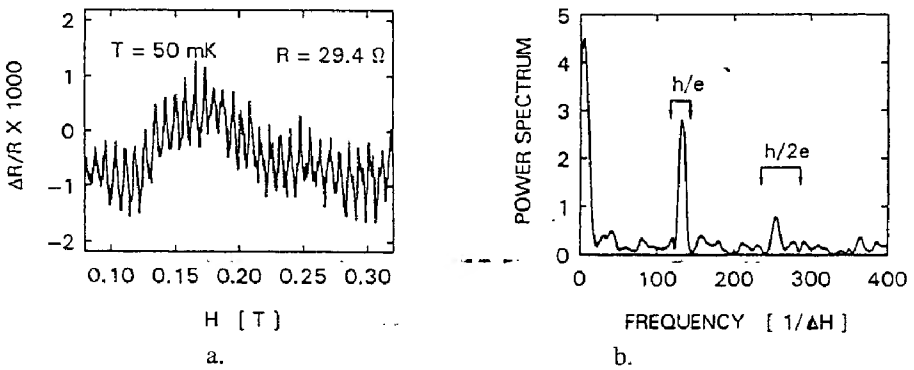


Figure 1.5: Experimental Aharonov-Bohm oscillations. a. resistance versus magnetic field; b. the Fourier transform of showing the h/e fundamental and h/2e first harmonic of the oscillatory signal of a.

fluxquantum added to the ring.
Now what happens experimentally. In figure 1.5 the resistance of a ring of the type shown in fig. 1.4 is shown in dependence of the applied magnetic field. Fig. 1.5a clearly shows the oscillatory pattern predicted by the "theory" presented before. As the pattern contains information of a periodic nature it can be beneficial to take the Fourier transform of the data in order to make this aspect even more clear. This is shown in fig 1.5b. Note that the largest peak, indicated by *h/e*, denotes the period associated with the addition of one fluxquantum per cycle of the resistance oscillation.

Note that in passing we have introduced a very important concept in mesoscopic physics. In "deriving" the oscillatory resistance (or likewise the conductance) of the system we have assumed that the *conductance scales with the probability for electrons to enter at one side and emerge from the other side*. This concept of transmission probability and its relation to the conductance will be found to be of central importance throughout the book.

1.1.d. *Single electron charging in a small island.*

The final example we want to discuss focuses on the effect of a *single electron* on the electrical conductance through a small metallic object. The effect relies on the Coulomb interaction due to the *electron charge* and for this reason it is denoted as Coulomb blockade. Figure 1.6a shows a typical set-up. A small metallic grain is embedded inside an *insulating* environment, e.g. an oxide. The whole structure is put onto a good metallic conductor, called the substrate. In addition a sharp tip is brought into close proximity to the grain, *without protruding the insulating film.* Some typical dimensions are a few nm's for the grain and < 1nm for the thickness of the insulator.
Classically, the insulating film surrounding the conducting grain forms a barrier which will prevent electrons to travel from the tip to the grain or from the grain to the substrate and vice versa. Quantum mechanically however we know that a particle can *tunnel* through this (classically forbidden) barrier, provided the barrier is not too thick. This

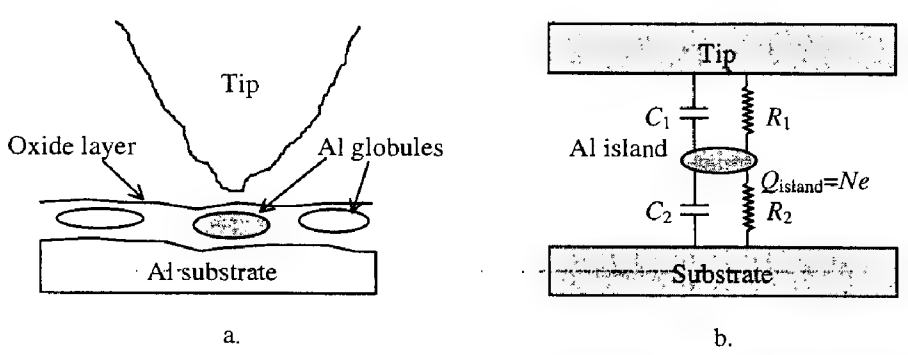


Figure 1.6. Coulomb blockade in a small metallic grain; a: Layout of the tip-grain-substrate; b. schematic circuit representation

dictates the thickness of the insulators. So, the barriers can be modelled as two *large valued* resistors, connecting the grain to the tip and to the substrate.
The configuration of figure 1.6a also includes *capacitances*. The grain will have a capacitance to the substrate as well as to the tip. In general the capacitance of a conductor put in an insulating environment decreases approximately linearly with the size. For the case of a metallic sphere of diameter d put in infinite free space:

$$C=2\pi\varepsilon_0 d \qquad (1.3)$$

In our case a rough estimate yields $C\sim10^{-17}$F. Combining both aspects allows us to represent the whole system schematically as shown in figure 1.6b.
What about the *number of free electrons* situated on the grain? In the limit of infinitely large resistances of the barriers the number of electrons will be classically determined, and simply be an integer: the electron either sits on the grain or it will reside elsewhere (tip or substrate). In the other limit of very small resistances the electron *waves* extend from the grain into the substrate or tip, making the number of electrons *actually sitting on the grain very badly defined.* So, if we assume the resistances R_1 and R_2 of the barriers to be large but finite, then charge transfer is still possible while the number of electrons at the grain is well-defined.
Now let us turn to the *energies* involved in the system. The first one in the electrostatic energy due to the charge Q on the grain, given by $E=Q^2/2C$, with $C=C_1+C_2$, i.e. the total capacitance of the grain. As we have seen, the number of electrons is an integer, or $Q=Ne$ (N=1, 2, 3, ..) and so the energy required to put the minimum amount of charge (i.e., one electron, or N=1) onto the grain equals

$$E_C=\frac{e^2}{2C} \qquad (1.4)$$

This quantity is called the *charging energy*.
The second energy involved is the thermal energy, kT, with k being the Boltzmann constant and T the temperature. Evidently if the electrons in the tip or substrate are shaken up by the temperature to such an extent that their energy is larger than the charging energy they may be able to get into the grain. So, to see effects due to the charging energy we have to require $E_c>>kT$. This translates into: low temperature and small grains.

Problem: Argue why a small grain size is required; estimate the temperature limit for our typical Al grain mentioned in the text.

Problem: Argue why *both* resistors R_1 and R_2 of the barriers have to be large.

Now we measure the total resistance of the structure by applying a (small) voltage V between tip and substrate and measure the current that results. As we have seen, the addition of one electron to the grain requires a *minimum* energy given by $-E_c$. So, as long as the voltage V across the structure is too small to provide electrons with an energy larger than the charging energy ($eV<\sim -E_c$) *no current can flow: the system is blocked.* This is called the *Coulomb blockade* due to single electrons. Increasing the voltage to beyond E_c/e will allow electrons to overcome the Coulomb blockade and current will start to flow. This is exactly what is shown in figure 1.7.

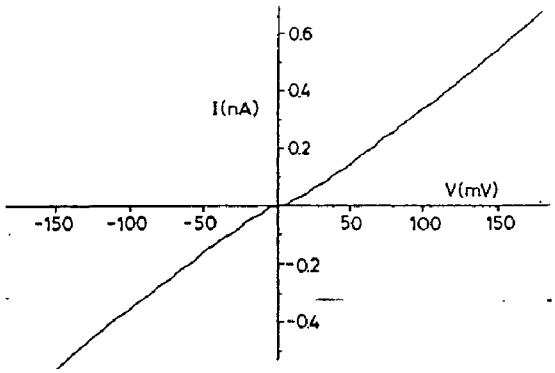


Figure 1.7. Experimental charging effects in a metallic grain

For voltages $|V|<\sim 8$ mV one clearly can see that no current flows. At larger voltages the current starts to flow rather abruptly. Note that at higher voltages more steps in the current are seen. Without providing detailed arguments for the moment, these steps are a consequence of a similar blockade, however now with 1, 2, 3, ... electrons sitting on the grain.

This experiment completes the short list of examples of mesoscopic phenomena. In the rest of the course these and others will be discussed in more detail.

1.2 Basic properties of conducting systems

In this section we briefly consider a few basic properties of free electrons in solids, most of which will be known already by the reader from introductory solid state physics and quantum mechanics courses. The reason for rediscussing these fundamentals is that we want to clarify a few relationships that commonly go rather unnoticed during the regular introductory courses. These focus onto the role of various characteristic length, time and energy scales in mesoscopic systems. Appreciating these, and understanding their mutual relations is of central importance throughout the rest of the course. Related to the length scales we will discuss the role of the dimensionality (3, 2, 1 or 0-dimensions) of mesoscopic structures, i.e. how do the (three) sizes of the sample system compare to the various characteristic length scales. Given its importance in electrical transport we also address the density-of-states (DOS) for electrons in solids, where the role of the dimensionality is evident.

1.2.a. *Length scales*

Let us first concentrate on the *length scales* that determine the way electrons move through a conductor.

= The most obvious one is the *size of the system*, L. Obviously the size can be different for the three directions in space, L_x, L_y and L_z. The specific length for each direction L_i (i=x, y, z) in comparison to a one of the other length scales introduced below determines the number of *relevant dimensions* or the *dimensionality* for the process related to this particular length scale; we will discuss this in more detail after the introduction of these other length scales.
= For the next length scale it is important to realise oneself that electrons behave as waves, with an energy dependent wavelength associated with it. With the electron of mass m having a kinetic energy E_{kin} and travelling at a velocity v its de Broglie wavelength λ follows straightforwardly from

$$E_{kin} = \frac{mv^2}{2} = \frac{p^2}{2m} \tag{1.5a}$$

and

$$p = mv = \hbar k = \frac{h}{\lambda} \tag{1.5b}$$

For the case of electrons in a solid one particular kinetic energy and wavelength comes into play. For this to appreciate we have to remember that electrons are particles with spin one-half. This implies that they are subject to the Pauli exclusion principle, which states that no two electrons can sit in the same available state of the solid. Consequently the electrons in the solid are "stacked" in the available states, the first electron put into the state of lowest kinetic energy (as usual in nature), the next one in the next higher state etc. With all electrons put in, the last electron will have the largest energy in the system, and this maximum kinetic energy is called the *Fermi energy* E_F. For this maximum energy E_{kin}=E_F eqs. (1.5a,b) immediately provide us with the associated Fermi velocity v_F, Fermi momentum p_F and *Fermi wave length* λ_F:

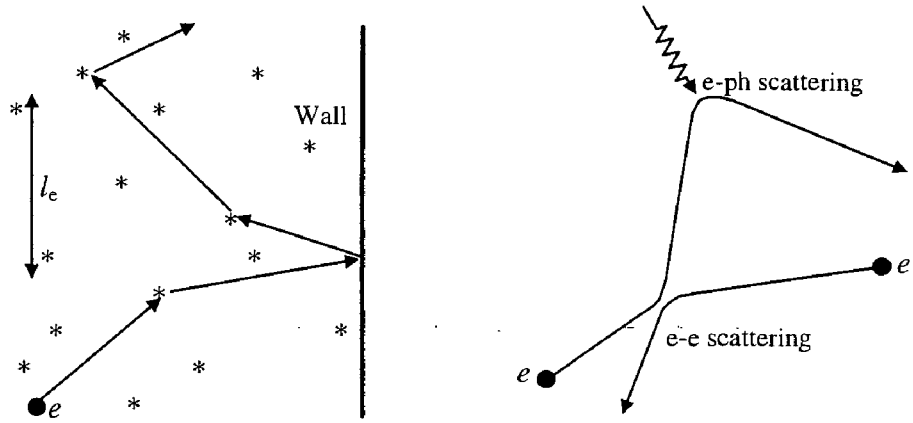
$$\lambda_F = h / \sqrt{2mE_F} \tag{1.6}$$

More details on E_F are given in the next subsection on energy scales.
For electron transport nearly always only those electrons with energy at or close to the Fermi energy are of importance (for an exception, see chapter 5, section 5.3 on persistent currents). Thus the Fermi wavelength is the relevant length scale in a variety of quantum effects, in particular those depending on wave-interference.

We can make a simple estimate of the Fermi wavelength by realising that the minimum wavelength λ_{min} that can be set-up (unambiguously!) in a system of electrons at an average distance d_{ee} will be approximately this value. As this minimum electron wavelength is associated with the largest kinetic energy, which is the Fermi energy, this implies that $\lambda_{min} \sim \lambda_F$, and so we immediately find

$$\lambda_{min} = \lambda_F \sim d_{ee} \tag{1.7}$$

= The next length scale we need to consider in electron transport relates to the fact that conductors need not be perfect. In textbooks on introductory solid state physics it is assumed that the solid forms an ideal regular crystal lattice, showing perfect similarity from unit cell to unit cell. "Real world" conductors may deviate from this ideal case and



a. b.

Figure 1.8. Scattering of electrons. a. shows elastic scatterings occurring at localised, time-independent scattering centres, like impurities or walls; b. shows inelastic scattering processes via time-dependent mechanisms such as lattice vibrations (or phonons) or electron-electron interactions.

show irregularities in the lattice (or more formally: absence of translation invariance). This breaks the fundamental condition for the electron wave to travel throughout the whole crystal, as any such irregularity will lead to (partial) *scattering* of the wave. This implies that the free length to travel, called the *mean free path*, no longer will be infinite. Scattering of the electron can result from different mechanisms thus leading to different mean free paths. The first to consider results from *static* faults in the crystal, like missing or displaced atoms (dislocations), foreign atoms (impurities) or walls (figure 1.8a). An electron incident on such fault will interact with it and become scattered. As the mass of this scattering centre, -typically an ion-, is very large compared to the electron mass virtually *no energy* will be exchanged with this "site", i.e. the scattering process is *elastic*. So the scattering can be characterised by $E_{initial}=E_{final}$ and $|k_{initial}|=|k_{final}|$; note, however, that the *direction* of the two initial and final vectors need not be the same: or $\vec{k}_{initial} \neq \vec{k}_{final}$. In case many such elastic scattering centres are incorporated in the solid we can define some average distance the electron can travel before it will experience an elastic scattering event: this distance is called the *elastic mean free path* (or *elastic scattering length*) l_e and it will become shorter with an increase of the density of scatterers. In the literature this length is also referred to as the *momentum relaxation length* l_m or l_{imp}. Note that also the boundary or edge of the system will effectively act as a (huge!) scatterer, as electrons will be fully reflected from this edge (they are not allowed to leave the crystal!). So, in small but otherwise ideal crystals, the effective elastic mean free path is of the size L of the system. Note that the average time between successive scattering events is given by $\tau_e=l_e/v_F$, the *elastic scattering time*.

The second influence acting on the crystal regularity is of a *non-stationary* or *time-dependent* nature. The most common example is the effect of temperature, which leads to lattice vibrations (in classical wording) or phonons (in quantum language)(figure 1.8b). An electron travelling through the crystal will experience the lattice distortion and become (partially) scattered by it: this leads to a reduction of the mean free path. In more quantum mechanical wordings: the electron travelling through the phonon sea (or phonon bath) responds to it via *electron-phonon interaction*. In contrast to the former elastic case the electron now picks up energy from the phonons: it even leads to an annihilation or absorption of a single quantum of lattice vibration or phonon. The exchange of energy implies an *inelastic* process, and the associated average distance travelled between "successive" inelastic scattering events is called the *inelastic mean free path* or *inelastic scattering length* l_i. Evidently neither the energy nor the wavevector of the electron under consideration are conserved in the scattering process. Given this energy exchange the characteristic length is also referred to in the literature as the *energy relaxation length* l_ε. In exactly the same way as before we can define an *inelastic scattering time* τ_i (or τ_ε) "between" scattering events: $\tau_i = l_i / v_F$.
For this specific case, i.e. inelastic scattering via electron-phonon interaction, the particular inelastic scattering length and time are represented by $l_{e\text{-}ph}$ and $\tau_{e\text{-}ph}$ respectively. It is of interest to note that at low temperature, given the rather small momentum carried by a phonon, the change in direction of the electron for each electron-phonon interaction event will be rather small as well. So electron-phonon scattering at low temperature leads to small-angle scattering.
Not only phonons may lead to inelastic interactions. Also the Coulomb interaction of the electrons among themselves leads to scattering with the exchange of energy. This process is called *electron-electron interaction*. The amount of energy exchange is strongly dependent on the initial difference in energy of the two electrons in the process, and so the associated scattering length $l_{e\text{-}e}$ and time $\tau_{e\text{-}e}$ will also show a strong energy dependence. It should be mentioned that for this case of interaction between particles of the same mass the scattering may lead to large changes in the direction of the k-vectors of the two electrons involved, even though only a small amount of energy is exchanged, i.e. it may lead to large-angle scattering.

= Now we are in the position to discuss a crucial difference with respect to the quantum mechanical consequences associated with the two types (elastic and inelastic) of scattering. It also will allow us to introduce one more length scale, the phase breaking (or phase coherence) length. This relates to the effect of the scattering on the *phase of the travelling electron wave*. To appreciate this difference let us consider a (probe) electron which is allowed to follow the same path twice. In case of elastic scattering the electron, retracing its path the second time, will experience *exactly the same environment* (impurities, walls/surfaces, voids) and so it will scatter in the same way as during its first travel. This predictability implies that also the phase of the scattered electron wave will evolve in a deterministic (i.e. non-statistical) way: with the electron starting at some position we can calculate exactly what the phase of the resulting wave will be at any other position in space, irrespective of how many elastic scatterings have occurred during its travel. So we conclude that *elastic scattering does not randomise the phase of the electron wave*.

This has to be contrasted to the case of inelastic scatterings. Now the details of each individual scattering event are unique. If the electron starts to do the path for the second time the scattering will be different, as it depends on mechanisms that act *statistically in time and space*. Because of the time-random nature of these scattering interactions the phase is not evolving in a predictable way but statistically, which leads to a blurring of the overall phase after a number of such inelastic events. So, in contrast to the elastic case, we find that *inelastic scattering randomises the phase of the electron wave.*
In conclusion, the two mechanisms leads to a fundamentally different behaviour of electrons contained within a restricted volume in a solid. While elastic mechanisms only *modify* the wave patterns of electrons contained in the system of interest which may lead to an increased complexity of the pattern, inelastic effects control the very *existence* of such patterns and ultimately may even completely suppress these.

This automatically brings us to the third inelastic scattering length scale, i.e. the *phase coherence (or phase breaking) length*, l_φ (with its associated time scale τ_φ). It represents the distance an electron can travel before its phase becomes randomised. From its very nature it will be clear that these phase breaking scales determine the degree to which the electron can experience processes where the phase is of importance, i.e. in quantum interference effects. As phase breaking and energy exchange are clearly related, we rather loosely intermix the use of the phase breaking length l_φ and the inelastic scattering length l_i (or l_ε) throughout the following chapters, although formally these are not identical.

= The last length scale we want to introduce emerges once a magnetic field is applied to our electron system. Such a field will interact with an electron travelling at velocity v via its charge e, leading to the Lorentz force $\vec{F} = e\vec{v} \wedge \vec{B}$, which is directed perpendicularly to the directions of the field and the velocity. The effect of the Lorentz force is that the straight zero-field path will become curved, resulting in circular *cyclotron orbits* (figure 1.9). Assuming for this moment that the electron is restricted to two dimensions, moving at the Fermi velocity v_F and that the magnetic field B is applied perpendicular to the 2D

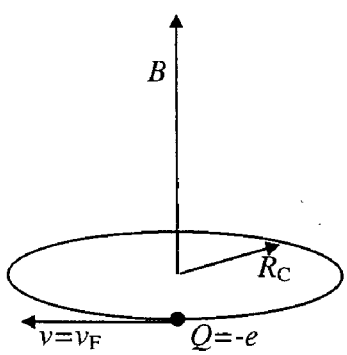


Figure 1.9. An electron moving in a magnetic field. The balance between the Lorentz force and the centrifugal force result in a circular or cyclotron motion.

plane, the resulting balance of forces yields

$$F_c = m^* v_F^2 / R_c = F_{Lorentz} = e v_F B \qquad (1.8a)$$

From this equation we can deduce the two important parameters, i.c. the cyclotron (circular) frequency

$$\omega_c = \frac{v_F}{R_c} = \frac{eB}{m^*} \qquad (1.8b)$$

and the cyclotron radius

$$R_c = \frac{v_F}{\omega_c} = \frac{m^* v_F}{eB} = \frac{\hbar k_F}{eB} \qquad (1.8c)$$

The electron mass m^* included in these equations does not need to be the free electron mass, m_0=9.31*10^{-31} kg. From introductory solid state physics we know that in the crystalline solid the relation between the kinetic energy and the velocity, the so called *E-k* or dispersion relation, is affected by the interaction of the electron wave with the crystal lattice. This leads to a modified or *effective electron mass* m^*. The effect on the mass is strongest for electrons at relatively small (kinetic) energies, i.e. close to the edge of an energy band. So, for simple metals with typically half-filled bands the effective mass commonly does not deviate too strongly from the free electron value m_0. In contrast in semiconductors with the electrons close to the conduction band edge (or holes close to the valance band edge) the effective mass may deviate strongly: in so called narrow gap semiconductors like InAs or InSb effective electron masses of ~0.03m_0 are found. In short, it is this mass which should be inserted in eqs. (1.8).
Note that if the electron is allowed to travel also in the third dimension (i.e. along the direction of the magnetic field) the Lorentz force in that direction will be zero, because of its vector-product nature. Thus the resulting trajectory will show a circular shape if projected onto the plane perpendicular to the *B*-field, while the movement in the third dimension will not be affected by the field.

1.2.b. *Energy scales*

In this subsection we will introduce a few *energy scales* that are of importance in mesoscopic systems. In particular the thermal energy kT, the Fermi energy E_F and the energy levels E_j due to confinement will be discussed. In addition a rather unknown energy scale will be discussed, called the Thouless energy, first introduced by Thouless.

= From subsection a. we know that electrons interact with the phonons in the system (the "phonon bath"). If the system is in thermodynamic equilibrium the density and energies of phonons are governed by the temperature T of the lattice. Strictly speaking, the only way the lattice "knows" its temperature is from the occurrence and distribution of its phonons! In particular the maximum energy of the phonons in the bath is given by the *thermal energy* $k_B T$, with k_B being the Boltzmann constant. For the sake of convenience the following approximate relation between the temperature and the associated energy can be kept in mind:

$$T=1\text{ K} \iff k_B T = 1.4*10^{-23}\text{ J} \sim 100\ \mu\text{eV}$$

= The second energy scale of interest follows from the fact that electrons obey Fermi-Dirac statistics. As discussed before this implies that no two electrons are allowed to occupy the same quantum state in a system, which makes electrons to become "stacked" in energy filling one state after the other. The maximum energy of the electrons (at zero temperature) is denoted by the *Fermi energy* E_F and is determined by the density of conduction electrons n_e. The well known Fermi-Dirac distribution $f_{FD}(E,T)$ is given by

$$f_{FD}(E,T) = \frac{1}{1+exp(\frac{E-E_F}{k_BT})} \qquad (1.9)$$

which is shown in figure 1.10. At zero temperature an abrupt transition from unity occupation for $E<E_F$ to zero occupation for $E>E_F$ is seen. For increasingly larger temperatures the transition becomes smeared over an energy interval of a few times the thermal energy k_BT, implying that at elevated temperatures empty states result below the Fermi energy while some states beyond E_F become occupied. In the majority of cases that we intend to consider we will assume low temperatures, i.e. a nearly discontinuous transition from unity to zero occupancy at E_F.

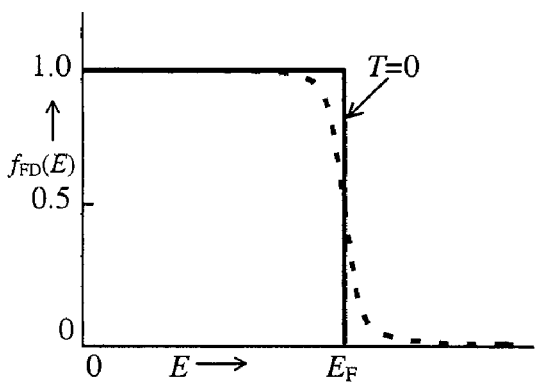


Figure 1.10. The Fermi-Dirac distribution function $f_{FD}(E)$ for $k_BT=0$ and $k_BT\sim0.1\ E_F$. Note the broadening of the transition around the Fermi energy given by a few times the thermal energy k_BT.

In order to evaluate the Fermi energy, we want to see its relation with the size of a system containing the electrons and the number of electrons it contains. This will also allow us to derive a few more fundamental aspects of electrons in the solid, such as the density-of-states (or DOS), and to discuss the effect of the dimensionality of the electron system. Let us first concentrate on the case that electrons are restricted to certain a range in one-dimensional (1D) space, i.e. a linear chain or wire. With the chain of length L we follow the common way to evaluate the electron states in this system by allowing only those electron waves whose wavelengths match the length (figure 1.11). As we assume that the electrons can not leave the 1D-wire the boundary condition for the system is zero amplitude of the wave(-function) at (and beyond) the ends of the wire; note that vanishing of the wavefunction outside L implies that the electrons are contained in a linear box of infinitely high and impenetrable walls.
The longest wave compatible with this condition forms a half wave spanning the length L, i.e. $\lambda_1 = 2L$: this forms the first or lowest (1D) eigenstate $|1>$. The second state results for one full wave of wavelength fitting the length, i.e. with a wavelength $\lambda_2 = 2L/2 = L$.

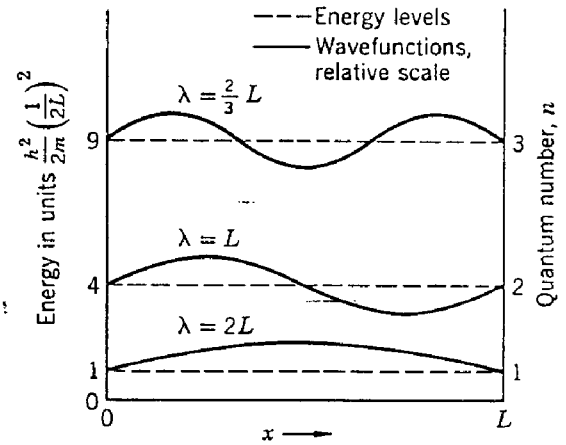


Figure 1.11. Energy levels of electrons inside a 1-dimensional box (or line). The successive states j=1, 2, 3, ... result from matching 1, 2, 3,.... half-waves within the length L of the box.

Similarly three half waves of wavelength $\lambda_3=2L/3$ for the third state, etc. can be set up, i.e. in general for the state $|j\rangle$

$$\lambda_j = 2L/j \qquad (j=1,2,3,...) \tag{1.10a}$$

For the associated k-vector of the state $|j\rangle$ we thus find

$$k_j = 2\pi/\lambda_j = j\pi/L \tag{1.10b}$$

and for the kinetic energies

$$E_j = \frac{\hbar^2}{2m}k_j^2 = \frac{\hbar^2}{2m}\frac{\pi^2}{L^2}j^2 \tag{1.10c}$$

It is of importance to note that the difference in energy between two successive levels

$$\Delta E_{j,j-1} \equiv E_j - E_{j-1} = \frac{\pi^2\hbar^2}{2m}\frac{2j-1}{L^2} \propto \frac{1}{L^2} \tag{1.11}$$

is larger if the length into which the electrons are confined is smaller, i.e. *strong confinement leads to large energy level splitting or separation.*

To see how we have to accommodate all the electrons available in the 1D-wire we need to see how many states there are within a certain range of k-values. From (1.10b) we see that the k-vectors are equally spaced at intervals $\Delta k=\pi/L$, and so we obtain for the *density-of-states* or *DOS* in 1D k-space

$$\rho_{1D}(k) \equiv \frac{dN_k}{dk} = \frac{1}{\Delta k} = \frac{L}{\pi} \tag{1.12}$$

Now assume that the wire contains N electrons in total, that are uniformly distributed along the chain yielding a *linear* electron density $n=N/L$ (per unit length, i.e. in units m^{-1}). To evaluate how many states are occupied by the N electrons we have to recognise that each k-state may contain more than one electron. We assume that the *degeneracy* is only 2-fold, with each k-state occupied by two electrons of opposite spin, denoted by the *spin-degeneracy* $g_s=2$. So, by integrating eq. (1.12) we immediately obtain the Fermi momentum k_F, which by definition equals the largest k value for any occupied state.

Thus,

$$N = \int_0^{k_F} g_s \rho_{1D}(k) dk = g_s \frac{L}{\pi} k_F \tag{1.13a}$$

or

$$k_F = N \frac{\pi}{g_s L} = \frac{N}{L} \frac{\pi}{g_s} = n \frac{\pi}{g_s} = n \frac{\pi}{2} \tag{1.13b}$$

This immediately yields the Fermi energy

$$E_F \equiv \frac{\hbar^2}{2m} k_F^2 = \frac{\hbar^2}{2m} (\frac{N\pi}{2L})^2 = \frac{\hbar^2}{2m} (n \frac{\pi}{2})^2 \tag{1.14}$$

In eq. (1.12) the (1D-) DOS in k-space, $\rho_{1D}(k)$, was derived. To calculate the *DOS in energy space*, $\rho_{1D}(E)$, we have to take into account the quadratic E-k relation, given by $k = \sqrt{(2m/\hbar^2)E}$ and so one finds

$$\rho_{1D}(E) \equiv \frac{dN}{dE} = \frac{dN}{dk} \frac{dk}{dE} = \rho_{1D}(k) * \sqrt{\frac{m}{2\hbar^2 E}} = \frac{g_s L}{\pi} \sqrt{\frac{m}{2\hbar^2}} \frac{1}{\sqrt{E}} \tag{1.15}$$

i.e. the 1-dimensional E-DOS follows an inverse square root dependence on energy.

In order to evaluate the Fermi momentum and the DOS in k- and E-space in 2 and 3 dimensions (2D and 3D) one has to follow the same line of reasoning as shown above. As the details can be found in the references [1] mentioned at the end of this chapter we only quote the results:

	1D	2D	3D	
k_F	$\frac{n\pi}{g_s}$	$\sqrt{\frac{4\pi n}{g_s}}$	$\sqrt[3]{\frac{6\pi^2 n}{g_s}}$	(1.16a)
$\rho(k)$	$\frac{g_s L}{\pi}$	$\frac{g_s L_x L_y}{4\pi^2} = \frac{g_s A}{4\pi^2}$	$\frac{g_s L_x L_y L_z}{8\pi^3} = \frac{g_s V}{8\pi^3}$	(1.16b)
$\rho(E)$	$\frac{g_s L}{\sqrt{E}} \frac{1}{\pi} \sqrt{\frac{m}{2\hbar^2}}$	$g_s A \frac{m}{2\pi\hbar^2}$	$g_s V \sqrt{E} \frac{m\sqrt{2m}}{2\pi^2\hbar^3}$	(1.16c)

Here n denotes the density of electrons taken for the appropriate "volume" in d-dimensional space (d=1, 2 or 3), i.e. in 2D per area A or m^{-2}, and in 3D per volume V or m^{-3}! Note that, in contrast to eq. (1.12), we have included the spin degeneracy factor also in the DOS in k-space to stress that each state can accommodate two spin directions. From the Fermi wave vector k_F in eq. (1.16a) we can derive the Fermi wavelength

$\lambda_F=2\pi/k_F$ ~ 2-3 times the interelectron distance d_{ee}, which is in accordance with our "statement" in eq. (1.7).

= As briefly discussed in the last example of the preceding subsection, the electronic charge of an electron implies that the positioning of an electron onto an isolated conductor (or island) requires a finite energy. This energy is a consequence of the finite capacitance of the island relative to its environment. This capacitance C scales (approximately) linearly with the size of the island, d. The electrostatic energy for the case a single electron is put onto the island is simply given by the well known expression $E_C=e^2/2C$ (see eq. 1.4), and is denoted the (single-electron) charging energy.

= We finalise this subsection by introducing an energy scale of a slightly less transparent nature (at least at first hand!). This energy scale, called the Thouless energy after the person who first introduced it in a clear way, is by no means difficult to derive. However, it is not so easy to demonstrate at this stage its importance in mesoscopic physics. Within this brief introduction we will not try to explain what impact it has on various phenomena. Only in future chapters we will show its importance.
The Thouless energy (denoted by E_C or E_{Th} (please do not confuse it with the charging energy, mentioned in section 1 of this chapter!!)), provides a measure of the relation between the energy of a state and phase evolution in a restricted area of space of size L. More precisely, it equals the shift of the eigenenergy at a change of the phase at the boundary of ~ π.
Assume an ideal (scattering free) system. A plane (electron) wave of wave vector k (and energy E) travelling over a distance L acquires a phase $\varphi=kL$. If the wavevector of the state is changed from k to $k'=\Delta k+k$ (with associated energy $E'=\Delta E+E$) the phase will change by

$$\Delta\varphi = \Delta k\, L \tag{1.17a}$$

From the quadratic E-k relation we obtain

$$\frac{dE}{dk} = \frac{\hbar^2}{m}k = \hbar\frac{\hbar k}{m} = \hbar v \tag{1.17b}$$

with v denoting the velocity of the wave. Thus, for the change in energy due to the change in k-vector by Δk one finds

$$\Delta E = \frac{dE}{dk}\Delta k = \hbar\Delta\varphi\frac{v}{L} = \hbar\Delta\varphi\frac{1}{\tau_L} \tag{1.17c}$$

with τ_L being the time it takes for the wave to travers the distance L. If we take $\Delta\varphi$~1, the *Thouless energy* is obtained as

$$E_{Th} = \frac{\hbar}{\tau_L} \tag{1.18}$$

Rephrased in words, the Thouless energy for a system of size L follows from the time it takes for an electron wave to cover this size.
The derivation given above was for the most simple case of a clean system, allowing plane waves to be used to represent the electron state. If the system is more complicated, and a state can not be characterised by a single wavelength (or k-vector), the Thouless energy can still be derived from the "time-of-flight" argument.

1.2.c. *Dimensionality*

The notion of *dimensionality* is of crucial importance in mesoscopic physics. In order to enhance our understanding we want to look somewhat closer on how we can establish a *d*D system for electrons, i.e. 0D (or d=0) for full confinement, 1D for transport in one direction, 2D for freedom within a plane and 3D for an unconfined system.
Let us confine the electrons in a certain region in 3D space, say in a box of size $\{L_x,L_y,L_z\}$. The "rigid confinement" inside the box implies that the wavefunctions for all the states are taken to be zero at the boundaries. For each direction i (=x,y,z) we can set up standing waves following eqs. (1.10), i.e.

$$\lambda_{i,j_i} = 2L_i / j_i \qquad (j_i=1,2,3,...) \qquad (1.19a)$$

with associated k-vectors

$$k_{i,j_i} = 2\pi / \lambda_{i,j_i} = j_i\pi / L_i \qquad (1.19b)$$

For the total kinetic energy of each level characterised by the set of quantum numbers $\{j_x j_y j_z\}$ we simply have to sum the three components

$$E = \sum_{j_i} E_{j_i} = \frac{\hbar^2}{2m}\sum_{j_i} k_{j_i}^2 = \frac{\pi^2\hbar^2}{2m}\left(\frac{j_x^2}{L_x^2}+\frac{j_y^2}{L_y^2}+\frac{j_z^2}{L_z^2}\right) \qquad (1.19c)$$

Now assume that we have a "box" such that $L_x >> L_y, L_z$, i.e. it is a needle-shape with the largest dimension along the x-direction. Consequently, the electrons can move in the x-direction much further than in the other two directions. From eq. (1.19c) it is obvious that the eigenenergies resulting from the small dimensions L_y and L_z are much larger than that those from the large dimension L_x (figure 1.12a). Evidently, this also holds for the differences in energies of the lowest (j_i=1) and second (j_i=2) level for the three directions: in the x-direction we find a rather narrow energy level spacing (figure 1.12b), while the other two directions y and z show much larger level splittings (figure 1.12c and d). The total energy level diagramme (figure 1.12e) demonstrates that, at low energies (i.e. the lower lying states of the system), only the level spacing resulting from the largest dimension is of importance. If the system is filled by putting electrons into the box we will do so starting from the lowest state $\{1,1,1\}$. If all electrons are accommodated (and so a certain electron density is reached, eq. (1.13a)) the energy of the topmost state by definition equals the Fermi energy. So, if the density is not too large, only states of the type $\{j_x,1,1\}$ will be occupied. The resulting electron system is said to be truly *1-dimensional*, with only the *lowest 1D subband being occupied*. If the electron density is increased we may reach the condition that it becomes energetically favourable to start filling also states like e.g. $\{j_x,2,1\}$, i.e. also the *second 1D subband* becomes occupied. So, a system is defined to be 1D whenever electrons occupy a range of states associated with one direction, with only the lowest state for the two other directions involved.

Problem: what condition should be met by the dimensions L_y and L_z to make that the first state in the second subband will be $\{1,2,1\}$.

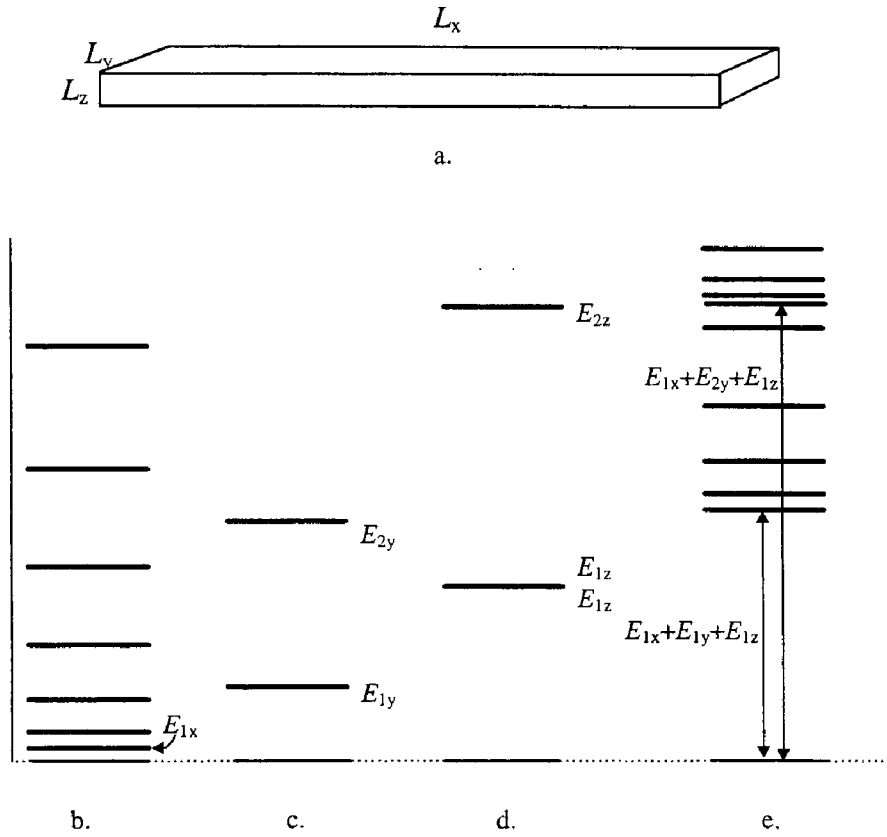


Figure 1.12. Energy levels of electrons contained in a long box of dimensions $L_x >> L_y$, L_z (top panel a.). Panels b., c. and d. show the level structure for the three orthogonal dimensions x, y and z. Note the much larger energy level spacings for the shorter sizes y and z, as compared to the long dimension x. The rightmost panel (e.) shows the total level diagram.

In a similar way the case of a 2D system can be discussed. Assume a box with two dimensions much larger than the third, e.g. L_x, $L_y >> L_z$. Electrons will now be allowed to occupy states characterised by the quantum numbers $\{j_x j_y, 1\}$, provided the electron density is not too large. The system is said to be truly *2-dimensional*, with only the *lowest 2D subband being occupied.* Increasing the density will, at a certain value, allow also states like $\{j_x j_y, 2\}$ to become filled, and so the second 2D subband becomes occupied. So, a 2D system is obtained when the electrons are kept in a "sheet-shaped" box and with an electron density such that only the lowest state in the third, smallest dimension is occupied.

If in either of the two preceding cases the electron density is increased such that not just one but a few subbands become filled, the system is said to be *quasi* 1D (or 2D). In the limit that the electron density is so large that many subbands (in all directions) become occupied the electron systems effectively becomes again 3-dimensional.

Before extending the notion of dimensionality beyond that associated with confinement alone it is useful to point onto an important implicit assumption made in our preceding discussion. For the very existence of k-vector quantisation in a certain direction (say x) it is crucial that the electron wave is allowed to travel back and forth a number of times along the length L_x, without losing its phase information appreciably. This implies that the phase coherence length should be considerably larger than the size of the system, or $l_\varphi >> L_x$. Note that this is an underlying (however crucial!) assumption which is nearly never mentioned in textbook derivations on the electronic energy states in solids!

This influence of a different lengthscale in the problem is a good example to generalise the concept of dimensionality. If a particular phenomenon is governed for instance by a particular type of lengthscale, say the phase coherence length l_φ, the ratio of this parameter to the size of the system determines the dimensionality *for that particular phenomenon*. So, if the system under consideration meets e.g. the condition $L_x << l_\varphi$ but $L_y, L_z >> l_\varphi$, than it is defined to be of a 2-dimensional nature for *phenomena governed by phase coherence*. Note that if in this case $L_x >> \lambda_F$, the system is completely 3D for effects that depend on the confinement.
Stated more generally, the dimensionality of a system can only be defined in relation to the phenomenon under consideration.

1.3 Transport regimes and material types

In the preceding section we have introduced the various length scales that govern electron transport in solids. Employing these results allows us in this short section to define the major regimes of electronic transport. In addition we will define the types of conducting systems more precisely and show how these can be realised in actual materials.

1.3.a. *Electronic transport regimes*

Electronic transport can occur in four different regimes. These are determined primarily by the lengthscales in the system (see section 1.2), i.e. the system size L, the elastic and inelastic scattering (or phase coherence or phase breaking) lengths l_e and l_i (or l_φ), and the Fermi wavelength λ_F. Just as a reminder we note that l_e is governed by the typical distance between impurities or other *static* lattice irregularities, while l_i (and by the same token l_φ) is determined by energy-exchanging type of interactions such as with other electrons or with phonons, i.e. by *non-stationary or time-varying* interactions. The Fermi wavelength is determined by the electron density (eqs. 1.13&1.14) and is roughly given by the average inter-electron distance (eq. 1.7).

To define the different regimes we first have to discriminate between the *quantum* case and the *classical* case. Evidently in the classical case we are "not interested" in any effect arising due to the *wave nature* of the electrons, and so the Fermi wavelength λ_F should be small compared to all other lengthscales in the problem. In contrast, for the quantum case the wave nature is a crucial ingredient, as this allows typical quantum effects like interference to occur. This also implies that the phase of the electron wave should be preserved sufficiently over the system, otherwise no "phase-memorisation" is possible.
A second way of discriminating different regimes can be set up along the lines of "how" the electrons follow their paths. If they are scattered very often during their traversal of the system they are called to behave *diffusively*: their paths are highly complex in space, similar to the "random walk" visible in the well known Brownian motion of a particle in a liquid (and, by the way, just as well demonstrated in the general behaviour in space of a "drunken sailor/student/..."). If, on the other hand, the electron can traverse the system without being scattered even once (i.e., it follows a straight line (in zero magnetic field)) its motion resembles that of a bullet in free space, and it is said to be *ballistic*.
Based on these arguments we can set up the following scheme.

Diffusive -----	Classical	$\lambda_F , l_i , l_e \ll L$
	Quantum	$\lambda_F , l_e \ll L , l_i$
Ballistic ------	Classical	$\lambda_F \ll L < l_i , l_e$
	Quantum	$\lambda_F , L < l_e < l_i$

In the following chapters we will discuss some of the major phenomena occurring in these different transport conditions or regimes.

b. *Metals and 2D semiconductors*

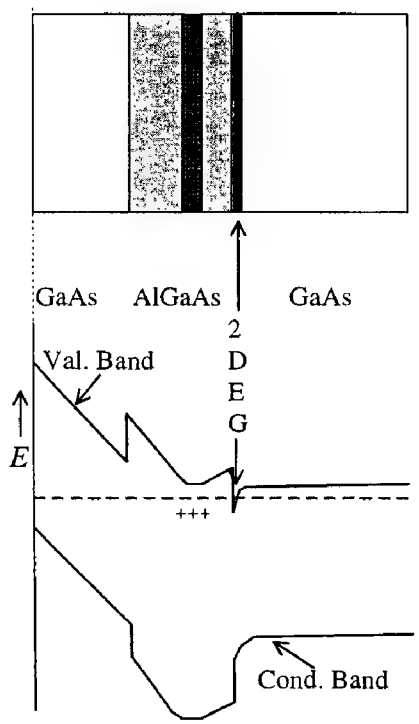
The second aspect we want to address in this section concerns the conditions we want to put onto our conducting systems. From introductory solid state physics (see references under [1]) we know that a conducting material like a metal or a semiconductor contains electrons that are "free" to move throughout the lattice. Their velocity (or kinetic energy) is determined by two parameters: the electron density n_e and the temperature.
Let us first concentrate on metals. In a metal each atom contributes at least one electron that becomes free (i.e. without being trapped by one atom/ion) to move throughout the whole piece of material: such a free electrons is *delocalised* and its wavefunction is non-zero "everywhere". The most characteristic feature of a metal is that the Fermi energy (i.e. the highest energy an electron can have at zero temperature and at equilibrium) is located inside one of its energy bands. For the simple metals this commonly means that approximately only (the lower) half of this band is filled. As we know the availability of empty, unoccupied states in the upper part of the band is a crucial feature for conduction to take place. Examples of simple metals are Na and K, while the nobel metals like Cu, Ag and Au are "close". The most characteristic feature of these metals is that all the filled

electron states are located inside a *sphere in k-space*, centred around k=0 and with a radius given by the Fermi momentum vector k_F. The spherical symmetry of the momentum distribution allows very significant simplifications in many calculations. Throughout this book we assume a single value for the Fermi wavevector and so we implicity assume that our conductors are of the simple nature as defined here.
As mentioned, in a metal *each atom* contributes at least one electron to the free electron "gas", and so the electron density n_e is comparable to the density of atoms in the lattice ($\sim 10^{29}/m^3$), i.e. n_e is large. Note that the Fermi wavelength, which is of the order of the inter-electron distance (eq. (1.7)), thus will be small, amounting to $\lambda_F \sim 2*10^{-10}$ or 0.2 nm.

Now let us turn to semiconductors. In a common intrinsic (=undoped) semiconductor, such as Si or Ge, at zero temperature all electrons will be bound to the atoms, and so the free electron density will be zero, implying the system will not show conduction. The system is an insulator with all electrons rigidly tied into the valence band. If, at increasing temperature, the thermal energy $k_B T$ becomes comparable to or larger than the typical binding energy (or the energy gap between the valence band and the conduction band), electrons can be released and put into the conduction band, where they can contribute to the conduction of the system. As many of the phenomena we are interested in require low temperatures (e.g., in order to preserve the phase of the travelling electron), it is evident that these *simple* semiconductors with zero electron density at zero temperature are not well suited to our purpose and so more sophisticated "semiconductors" are needed.
The electronic properties of semiconductors strongly depend on the degree of *doping*. Doping means the inclusion of a small concentration of appropriately selected foreign atoms in the semiconductor. Provided the density of these "Fremdkoerper" atoms is not too small they may release one electron even at zero temperature, generating a non-zero free electron density. As the electron density now is determined by the density of doping atoms, which is only a small fraction of the total number of atoms, the electron density will be much smaller than for a metal (typically $n_e \sim < 10^{25}/m^3$, i.e. a Fermi wavelength of ~ 10 nm). Note that also in this case the distribution of momenta in k-space will be spherical, complying with our simplifying condition. So, the resulting electron system behaves as a metal, however with a small electron density.
It should be noted that such a low-electron-density metals (which, for the sake of simplicity and in line with the common use in the literature, we will continue to call semiconductor!) have one highly attractive property, namely, they allow a full control of the electron density. By mounting a metal plate in close proximity to the low-density electron system (e.g. a 2D layer) and applying a voltage to this so-called *gate*, it electrostatically affects the charge density. In particular a negative gate voltage pushes electrons away from the vicinity of the gate electrode, in this way locally reducing the electron density in the 2D semiconductor system. This electrostatic controlling effect is employed in semiconductor devices such as field effect transistors or FET's. These FET's are the most important semiconductor devices used in chips (IC's) employed in PC's. It should be noted that this can not (yet!) be done in a metal, because of the very high electron density.

From the preceding section 1.3.a. it is clear that, in order to bring the structures into the various regimes (classical<=> quantum; ballistic<=>diffusive), these have to meet rather

stringent requirements in terms of the ratio between length scales. From appendix A we know that, employing modern electron beam lithographic techniques, structures can be made with sizes L<~ 30 nm. Comparing this minimum feature size to the Fermi wavelength of metals (a few tenths of nm) and of semiconductors (a few tens of nm) it will be clear that semiconductor systems more readily will allow us to enter the quantum regime than metals will do.
Here we want to introduce the most important semiconductor structure (at least for our purpose), the *2-dimensional electron gas* or *2DEG*. It can be realised in various semiconductor structures, but we will exclusively concentrate on so called *heterostructures*. Semiconductor heterostructures are grown to include a number of different (=> hetero)-layers, which allows the use of two crucial ingredients: a thorough confinement of the charge carriers in one direction (i.e. the growth direction) and the spatial separation of the doping atoms from their "offspring-" electrons. If this growth is performed with great care the resulting crystalline structure is very clean and nearly free of defects. Figure 1.13 shows the basic features of such a system, in this case built from GaAs and GaAlAs as the two different types of materials for the layers. Figure 1.13a shows the actual layer structure. Note that the doping atoms (+++) are inserted in the GaAlAs layer above the GaAs layer, at a certain distance away from the interface between the two. Figure 1.13b shows the *energy band diagram* of the structure. The most



a.

b.

Figure 1.13. A GaAs-GaAlAs heterostructure containing a 2-dimensional electron gas of very large elastic mean free path or mobility; surface is left, inner part of the structure is right.

a. depicts the grown structure with the different layers; note the AlGaAs-layer containing the n-type dopants (darkest area) that deliver the electrons forming the 2DEG at the lower-AlGaAs-GaAs (= rightmost) interface.

b. shows the energy diagram, depicting the depth- (or z-)dependent position in energy space of the edge of the conduction band (CB) and the valence band (VB); note the Fermi energy E_F in the system.

noteworthy detail in the diagram is the step in the (conduction- and valence-) band-edges occurring when changing from one material to the other, commonly called the *band-edge discontinuity*. This step at the interface can be employed to strongly confine the electrons in the growth- or z-direction. The (assumed) n-type doping atoms have released their electrons, leaving them as positively charged ions (indicated by +'s in fig 1.13b); the electrons assemble at the GaAs side of the lower interface. Note that the Fermi energy settles itself such as to cut through the steep, triangularly shaped notch in the conduction band edge, near the lower AlGaAs-GaAs interface. The electrons are free to move in the two perpendicular x- and y-directions, thus forming the 2DEG.
As mentioned before the crystalline structure can be grown nearly free of defects. This implies that electrons moving (in 2D) along the interface will experience virtually no elastic scattering and so their elastic mean free path l_e will be very large. Here the additional trick employed in such heterostructures comes to work. It is found that l_e is limited mostly by the ionised donor atoms that were left behind in the GaAlAs layer. As the donors are rather far away from the electrons (reason to call them *remote dopants*) the electrons only experience the weak potentials left from the unity-charged ions: they are said to interact with the "Coulomb tail" of the charged impurity.
State-of-the-art grown layers may show elastic free paths for the 2D electrons up to 100mm or more. Note that this is ~ 3000 times as large as the typical Fermi wavelength and the minimum feature size that can be fabricated. This evidently paves the way to use such electron systems in a variety of ballistic and quantum experiments. In addition these structures can be used as the starting system for further reduction of the dimensionality to 1D or 0D, as we will see in chapters 3 and 4.

<u>General references to the subject</u>

1. Introductory level textbook on Solid State Physics, e.g.
"Introduction to Solid State Physics", 7th edition, C. Kittel (John Wiley & Sons, New York, 1996), chapters 6, 7 and 8;
"Solid State Physics", H. Ibach and H. Lüth (Springer-Verlag, Berlin, 1993), chapters 6, 8, 9 and 12.

2. "Quantum Transport in Semiconductor Nanostructures" by C.W.J. Beenakker and H. van Houten, in Solid State Physics, ed. H. Ehrenreich & D. Turnbull, Vol 44 (Academic Press Inc., Boston, 1991), sections 1-4

3. "Electronic Transport in Mesoscopic Systems", S. Datta (Cambridge University Press, Cambridge (UK), 1995)

2. Diffusive electronic transport: classical and quantum regime

Keywords: Drude model, Einstein relation, conductance quantum, classical size effects, coherent backscattering, weak localisation, UCF

2.1 Introduction

At room temperature electrons travelling in a metallic solid conductor do so in a very erratic way. In stead of traversing a wire in a straight line, they in contrast only arrive at the output side after experiencing many collisions or scatterings throuhout the wire. As introduced in chapter 1 (section 3) the "straight line" regime is called ballistic, while the "intense scattering" case is denoted as the diffusive regime. In addition, we have to take into account the role of the wave nature of the electrons. If the scatterings are such that the phase is strongly destroyed over very short time intervals (or short distances) the regime is called classical, while the opposite limit is the quantum diffusive regime. This implies that, commonly speaking, the quantum regime will be accessible only in small structures and at low temperatures.
This chapter describes the transport of electrons in solids in the diffusive regime. Despite its familiarity a short review of the classical limit is useful. This provides a smoother transition to the inclusion of quantum mechanical effects on the conductance. In addition it allows us to clearly identify the differences in comparison to the second main class of transport, i.e. the ballistic regime (see chapters 3 and 4).
From section 1.2 we know that diffusive transport is characterised by the requirement for the electron to be scattered many times while traversing the system of size L. This implies that in the diffusive regime we require the condition $l \ll L$ to hold for the mean free path. To what extent quantum effects are of importance is governed by the inelastic length (or the phase coherence length l_φ, as we continue to loosely assume these two to be the same (see subsection 1.2.a)) relative to the system size. As presented in chapter 1 the two regimes (classical and quantum) are defined by comparing the various lengthscales in the system:

	!	Classical	$\lambda_F, l_i, l_e \ll L$
Diffusive	-------- !		
	!	Quantum	$\lambda_F, l_e \ll L, l_i$

Section 2.2 contains the concise review of the classical diffusive case. Although some aspects needed for our discussion can be found in common textbooks, the level is not always appropriate for our approach. In particular the discussion concerning the diffusion coefficient and the Einstein relation may be new for various readers. In addition it discusses the effect of the size of the system on the transport properties, including the properties of the walls/edges and the influence of an applied magnetic field. These phenomena go under the name of classical size effects. The following section 2.3 concentrates on quantum effects in the diffusive regime. Although this is nearly

exclusively limited to one effect, i.e. the so-called weak localisation, its importance for mesoscopic phenomena is so large that it deserves this position. In particular, it allows the introduction of a number of concepts, -such as the notion of transmission of (electron) waves to evaluate the conductance and the effect of the magnetic field on the wave properties of electrons-, which are of great use throughout the book as a whole. In addition, connection will be made to its optical analogue, -weak localisation of photons-, an effect also found in acoustic wave-fields.

2.2 Classical diffusive transport and classical size effects

In this section we will briefly reconsider the classical electron transport as described by the Drude model. As discussed before the classical diffusive regime implies that the relation between the various length scales in the system is defined as

$$\lambda_F , l_i , l_e << L$$

The main aim is to introduce a few quantities used to characterise conduction, such as the mobility and the conductivity, and to encounter for the first time the "unit of conductance" e^2/h. In addition we will introduce the Einstein relation for the conductivity which provides a neat and convenient bridge from the classical diffusive case to the quantum case of the next section. Finally we will discuss what *classical* effects arise when confining the electrons into a narrow channel, in particular concentrating on the nature of the scattering (diffusive or specular) at the walls of the channel.

2.2.a. *Diffusive transport: some fundamentals*

Drude's model, which was developed in the pre-quantum area at the beginning of the 20th century, is based on the assumption that the electron system can be described as an ideal gas. A particle of the gas can move freely, only limited by occasional scattering events (e.g. via mutual collisions and/or scattering onto local "disturbances"). This scattering, characterised by the average time interval between scattering events or the scattering time τ, is assumed to be of such a nature that the particle looses *all information about its former conditions*, such as *kinetic energy, velocity and direction*. Now we have to realise that in equilibrium the electrons from the ensemble move randomly, but uniformly distributed over all directions, and so the *mean velocity* $<\vec{v}>$ of the whole ensemble (taking into account all directions) evidently equals zero. If an electric field $\vec{E}$ is applied, the electrons are accelerated by the resulting force $-e\vec{E}$. The mean velocity now depends linearly on time

$$<\vec{v}(t)>= -e\vec{E}t / m \qquad (2.1)$$

with m denoting the electron mass. As, on average, after a time τ the electrons will become scattered, the resulting average velocity of the ensemble (or drift velocity) is given by

$$\vec{v}_d \equiv <\vec{v}(\tau)>= -\frac{e}{m}\vec{E}\tau \equiv -\mu\vec{E} \qquad (2.2)$$

which defines the *mobility* of the electrons, $\mu=e\tau/m$.

For the conductivity of the system we find

$$\sigma = \frac{\vec{j}}{\vec{E}} = \frac{-en_e < \vec{v} >}{\vec{E}} = \frac{e^2 n_e \tau}{m} \qquad (2.3)$$

with j denoting the current density. This is the famous *Drude expression* for the conductivity.

Now we want to proceed one step further, adding some quantum building blocks to this classical foundation. From chapter 1 we know that the electron density can be related to the Fermi wavevector k_F, the precise relation being determined by the dimensionality of the electron system (see eq. 1.16a). In addition we know that only those electrons with energies close to the Fermi energy can contribute to the current, which is a consequence of the requirement to have unoccupied states available immediately above the equilibrium energy of the particles, as resulting from Pauli's exclusion principle; below we will return to this aspect. This allows us to relate the scattering time and the mean free path in a very simple way.

Considering the 2D case we simply can write

$$n_e = \frac{g_s}{4\pi} k_F^2 \qquad (2.4a)$$

with g_s denoting the spin degeneracy factor (g_s=2). For the mean free path we know

$$l = v_F \tau = \frac{\hbar k_F}{m} \tau \qquad (2.4b)$$

Combining eqs. (2.3) and (2.4) we find

$$\sigma_{2D} = g_s \frac{e^2}{h} \frac{k_F l}{2} = \frac{2e^2}{h} \frac{\pi l}{\lambda_F} \approx \frac{2e^2}{h} \frac{l}{d_{e,e}} \qquad (2.5)$$

with the last approximate equality derived from eq. (1.7).

Here we encounter for the first time the famous "*natural unit of conductance*" e^2/h, a quantity which we will come across very often in the following chapters and which we will find of central importance in the description of the conduction in solids.

Considering eq. (2.5) somewhat closer, it relates the conductivity (in 2D) to the ratio of the free travel length and the inter-electron distance (or the Fermi wavelength). Intuitively one will feel that, if an electron is so heavily scattered that it is not capable to run freely for at least one wavelength, this will strongly affect the nature of the conductance. Putting it a little further, such strong scattering may effectively lead to the inhibition of the movement of the electron, i.e. the system will tend towards becoming some type of *insulator*. If, on the contrary, the free path is considerably larger than the Fermi wavelength, the system behaves as a regular metal.

So we may (indeed reluctantly, as far as the handwaving arguments given above, but this can be argued in a much stronger way!) conclude that the unit of conductance approximately separates two fundamentally different conduction regimes defining the boundary for such a *metal-insulator* transition in a solid.

In the discussion just preceding eq. (2.4) we mentioned the essential role of free states with energies immediately above (and below) the Fermi energy E_F. This can be made clearer by rewriting eq. (2.5) in such a way that the availability of states, quantified by the

density-of-states or DOS at the Fermi energy, $\rho(E_F)$, is directly visible. The resulting expression can be generalised to all dimensions and is called the Einstein relation. Equation (2.5) was obtained for the 2D case, in particular via eq. (2.4a) for the Fermi wavevector. From chapter 1 (eq. 1.16c) we know that the DOS for the 2D case per unit area is given by

$$\rho_{2D}(E) = \frac{g_s m}{2\pi\hbar^2} \tag{2.6}$$

This allows us to rewrite expression (2.5) as follows

$$\sigma_{2D} = \frac{g_s e^2}{h}\frac{1}{2}k_F l = e^2\frac{g_s}{2\pi\hbar}\left(\frac{m}{\hbar}\frac{\hbar k_F}{2m}\right)l = e^2\frac{g_s m}{2\pi\hbar^2}v_F l/2 \tag{2.7a}$$

or

$$\sigma_{2D} = e^2\rho_{2D}D_{2D} \tag{2.7b}$$

The quantity $D_{2D}=v_F l/2$ is called the 2D *diffusion coefficient*. Expression (2.7b) which we derived for the two-dimensional case, can be shown to hold in all dimensions and so it can be generally written as

$$\sigma = e^2\rho(E_F)D \tag{2.8}$$

which is the well known *Einstein relation* for the *conductivity*. It formally expresses the role of the DOS at the Fermi energy for the conductivity. In particular it demonstrates that a non-zero conductivity requires the density-of-states at the Fermi energy to be non-zero as well.

The diffusion coefficient D depends on the dimensionality d (=1, 2 or 3), following

$$D_{dD} = \frac{1}{d}v_F l = \frac{1}{d}v_F^2\tau \tag{2.9}$$

It expresses in which way scattering affects the average velocity of the particle in time. More precisely, it quantifies to what extent scattering maintains a relation between the velocity of an individual particle before and after the scattering event, i.e. it formally defines the (degree of) *velocity-velocity correlation*. This is what is clearly seen in the expression for D due to Kubo

$$D = \int_0^\infty \langle \vec{v}_x(t)\vec{v}_x(0)\rangle dt \tag{2.10}$$

Here <...> denotes the average over all possible directions in velocity- or k-space which the initial velocity $v(0)$ can take; the subscript x denotes the direction of the applied E-field. Via this spatial integration/averaging the dimensionality d enters into eq. (2.9). Although we will not derive the Kubo expression (2.10), as an example we evaluate the diffusion coefficient from it for the 1D case (d=1), assuming the Drude condition, i.e. complete loss of memory after each scattering event. Assuming an avarge scattering time interval τ it implies that for $t<\tau$ the inner product of the two vectors $v_x(t)$ and $v_x(0)$ will be simply $(v_x(0))^2$. For times $t>\tau$ the product will become random, both as far as the time integral as well as the k-space averaging is concerned, i.e. the total summed contribution will be zero. As in 1D the velocity only has two possible directions (+x and -x: the k-

space is restricted to a line!) this yields $(v_x(0))^2=v_F^2$, i.e. from eq. (2.10) we find $D_{1D}=v_F^2\tau$, which is in agreement with eq. (2.9).

Problem: during the time τ the velocity changes due to the applied electric field. Argue that in general the effect of this acceleration will be negligible in evaluating the diffusion coefficient. Quantify the conditions for this to hold.

From the Kubo expression (2.10) it is evident that if scattering does *not* destroy the correlation between the initial and final state of the electron, this will affect the value of the diffusion coefficient and so the conductance via the Einstein expression. In the following section 2.3 we will see that quantum coherence leads to such deviations from the simple expression (2.9).

In order to gain a more general understanding of the diffusive motion of an electron let us see how, in a purely classical approximation, the position of such a particle subject to random scattering events evolves in time. Figure 2.1a shows how such a particle will behave in space and time. Schematically figure 2.1a illustrates how a densely peaked distribution with all particles at $r=0$ at $t=0$ evolves in time, with the distribution becoming increasingly broader at later times (t_1 and t_2).
The equation governing the classical diffusion of particles is simply given as

$$\frac{\partial n(\vec{r},t)}{\partial t} = D\nabla^2 n(\vec{r},t) \qquad (2.11)$$

Limiting ourselves for simplicity to the 2D case the solution (obtained by introducing

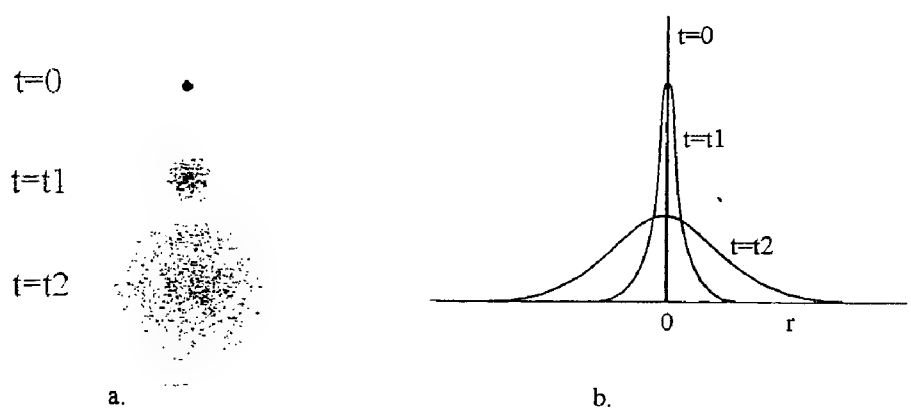


Figure 2.1. The diffusive motion of a classical particle in space in dependence of time. a. shows in real space how a densely peaked distribution of particles at r=0 at t=0 becomes increasingly diluted in time (t_1 and $t_2>t_1$). b. shows the Gaussian shape of the probability density n(r,t) to find a particle at position r at time t.

cylindrical coordinates) for the distribution is

$$n_{2D}(\vec{r},t) = \frac{n_0}{4\pi Dt}\exp(-\frac{r^2}{4Dt}) \qquad (2.12)$$

Note that this expression will be different for the case of other dimensionalities (1D or 3D). From eq. (2.12) we also can calculate the mean distance the ensemble of particles diffuses away from the origin at a given time t. After some mathematics one finds

$$l(t) \equiv <|\vec{r}(t)|> \equiv \frac{\int |\vec{r}| n(\vec{r},t) d\vec{r}}{\int n(\vec{r},t) d\vec{r}} = \sqrt{Dt} \qquad (2.13)$$

This implies that, in the *diffusive regime, the average displacement follows a square root behaviour versus time.*

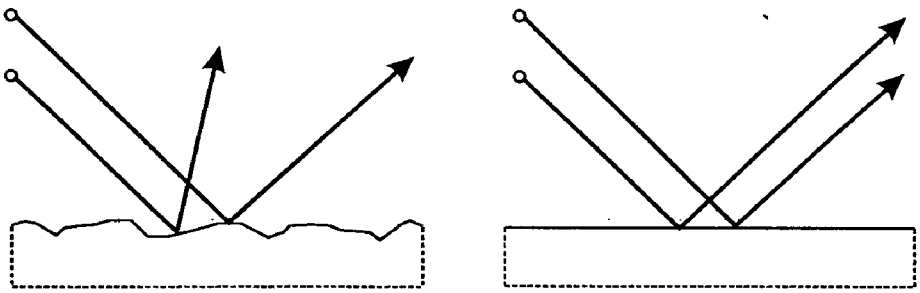
Expression (2.13) allows the straightforward evaluation of the phase-coherence length and the inelastic length in the diffusive regime in terms of their associated times τ_φ and τ_i respectively

$$l_{\varphi,i} = \sqrt{D\tau_{\varphi,i}} \qquad (2.14)$$

This relations will be employed at various occasions in the following sections and chapters.

2.2.b. *Classical size effects*

In the discussion of the various length scales in chapter 1 we have seen that the boundaries of the system may have an effect on the transport properties, as they act as elastic scatterers. We now want to investigate this a little more systematically. The effects resulting from the influence of the boundary go under the name *classical size effects*. The wall or boundary of the system will reflect electrons impinging on it from the interior or *bulk* of the system. This reflection can be of two types (figure 2.2). Either the incoming electrons are scattered and reflected randomly in all directions in (half-)space, yielding an isotropic distribution of the outgoing particle flux in the ideal case (see fig 2.2a). In the



a. b.

Figure 2.2. Electron reflection at the wall or boundary of the sample. a shows diffusive reflection with the incoming electrons being reflected isotropically in all directions; b the incoming electrons are specularly reflected, with the wall acting as a flat mirror

other limit the particles are specularly reflected at the wall, with the boundary in this way acting as a perfect mirror (figure 2.2b). Note that specular reflection implies that the *momentum of the particle is conserved along the wall*, while it is reversed perpendicular to it (in this way conserving energy, as required for an elastic process!).
What conditions has the wall to meet to provide specular reflection? Realising that our classical particles also have a wave nature, we can argue simply in analogy to the optical equivalent. This implies that whenever the wall is "wavy" on a length scale λ_F the electron wave diffracts and will become dispersed over different directions. This quantum mechanical probability distribution over angle has the classical equivalent that the electrons become distributed over a range of reflection angles and so specularity is affected. For metals, with $\lambda_F \sim 0.2$ nm, this implies that the wall has to be atomically flat, a very severe condition for any structure that can be made artificially (see Appendix A). In practice this means that specular reflection at a metal surface nearly never occurs. In contrast for semiconductors the typical Fermi wavelength is 10-50 nm. This is in reach of present day nano-fabrication techniques, and we will see that specular reflection indeed is found.

We now discuss the influence of wall scattering, including the role of an applied magnetic field. We can distinguish four cases, i.e. specular and diffusive scattering at zero and non-zero magnetic field.

I. Zero magnetic field, B=0

a. Specular wall reflection
Figure 2.3a shows the effect of the wall for this case. Because of the mirror-like behaviour of the wall the momentum of a particle hitting the wall will be affected in a fully predictable way. The longitudinal component (along the x-direction of the wall) p_x will be conserved, or $p_{x,in}=p_{x,out}$, while for the perpendicular component we immediately find $p_{y,in}=-p_{y,out}$. As we are interested in the current flow in the x-direction along the wall this implies that with the average momentum $\langle p_x \rangle$ also the average velocity $\langle v_x \rangle$ is

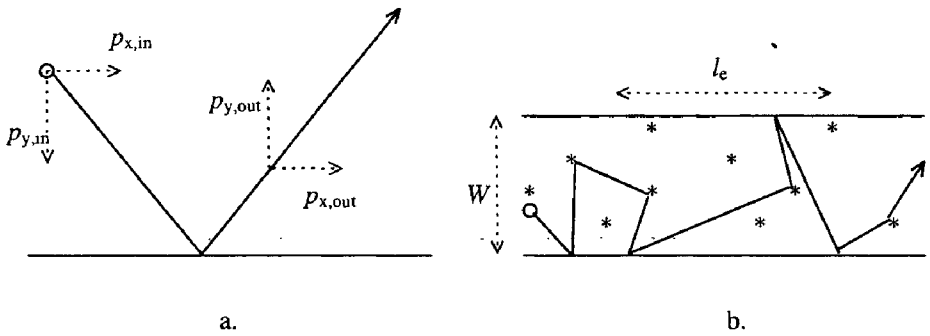


Figure 2.3. The effect of specular (a.) and diffusive (b.) reflection at the wall on the transport of electrons.

conserved. So, from the Drude expression eq. (2.3) we immediate conclude that the conductivity is not affected by specular reflection at the boundary of the system.

b. Diffusive wall reflection.

In this case a particle incident on the wall may become scattered randomly in all directions (figure 2.3b). This implies that the wall effectively acts as "just another scatterer" being added to the system. In order to evaluate its effect more quantitatively we introduce a second wall, in parallel to the first at a distance W, thus forming a channel (note that for simplicity we limit the discussion to a 2D system). The two walls that act as diffusive scatterers effectively lead to an increase in the density of scatterers. This results in a reduction of the mean free path approximately given as

$$\frac{1}{l_{eff}} \approx \frac{1}{l} + \frac{1}{W} \qquad (2.15a)$$

yielding an effective scattering time $\tau_{eff}=l_{eff}/v_F$, and the (Drude) conductivity becomes approximately

$$\rho(W) \approx \frac{mv_F}{e^2 n_e l_{eff}} = \frac{mv_F}{e^2 n_e l}\left(\frac{l+W}{W}\right) = \rho_0\left(\frac{l+W}{W}\right) \qquad (2.15b)$$

where ρ_0 denotes the infinite-size or bulk Drude conductivity. A more accurate expression can be found in ref. 2. Two limits can be distinguished. If the width is much smaller than the mean free path eq. (2.15b) yields

$$\rho(W) \sim \rho_0 \frac{l}{W} \qquad (W \ll l) \qquad (2.16a)$$

For the 2D case the more accurate calculations shows that a (relatively weak) logarithmic dependence needs to be added by multiplying eq. (2.16a) by $1/\ln(l/W)$.
In the opposite limit $l \ll W$ eq. (2.15b) yields a weak dependence on the width following

$$\rho(W) \sim \rho_0\left(1+\frac{l}{W}\right) \qquad (l \ll W) \qquad (2.16b)$$

From the preceding discussion we can conclude that the nature of the reflection at the wall can be derived from a systematic experimental study of the width-dependence of the resistivity $\rho(W)$.

II. Non-zero magnetic field (low field regime): magneto transport.

In chapter 1 we have seen how a magnetic field affects the motion of an electron due to the Lorentz force acting via the velocity and the charge of the particle. In the plane perpendicular to the B-field the resulting motion is circular with a radius given by eq. (1.8c) as $R_c=mv_F/eB$, the cyclotron radius. The question is how this new lengthscale will affect the transport in case the size of the system is reduced by the introduction of walls.

a. Specular wall reflection

In the preceding case of zero magnetic field we arrived at the conclusion that the longitudinal transport was not affected by the wall in case of ideal specular reflection. The basic argument leading to this conclusion was that the longitudinal component of the momentum is not affected by the mirror-like scattering. The same argument continues to

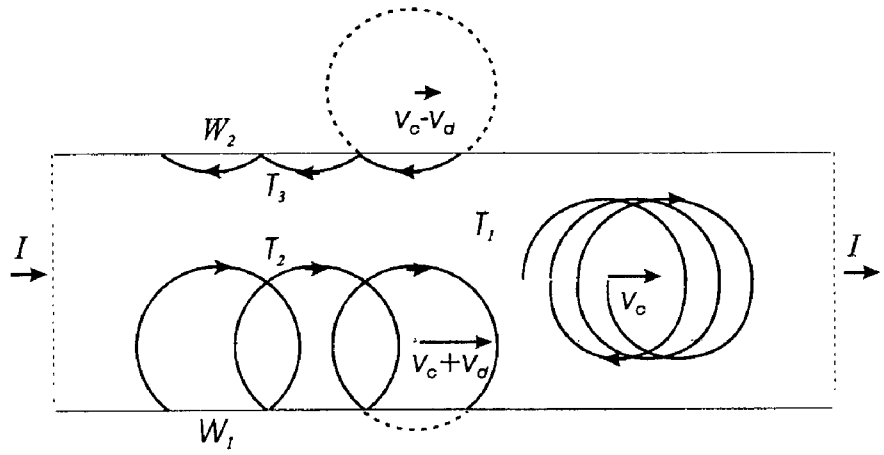


Figure 2.4. Electrons moving in a magnetic field in a 2D channel and between two specularly reflecting walls W_1 and W_2. The magnetic field is taken perpendicular to the plane of the picture. The "central" trajectory T_1 is the normal bulk cycloidal path formed by the sum of the electric field induced longitudinal velocity v_c and the cyclotron motion due to the magnetic field. The two trajectories T_2 at wall W_1 and T_3 at W_2 are each others complement as far as wall scattering is concerned; note that these paths tends to follow the wall in a peculiar sequence of bounces and circle segments, called skipping orbits.

hold for the non-zero *B*-field case, although this is not as self-evident as it may seem at first sight. In order to appreciate this one should note that the Lorentz force, leading to the formation of cylotron orbits (see eqs. 1.8), induces a coupling of the two (orthogonal) components of the velocity in the plane perpendicular to the magnetic field, a coupling which is absent at zero field. So a change of the component perpendicular to the wall is accompanied by a modification of the longitudinal component as well. We will find, however, that the *combined* influence of the two wall leads to a zero *net* effect.
Figure 2.4 shows schematically what happens. Let us assume that the field is such that the distance *W* between the walls is larger than the cyclotron diameter. Any path that keeps the central or bulk part of the channel, e.g. trajectory T_1 in figure 2.4, will not feel the wall. Thus, with its transport properties not being affected, the current carried by the centre part is also not affected.
Now consider an orbit that may hit the wall, such as T_2 in figure 2.4. Because of the specular reflection (at wall W_1) the resulting orbit forms a sequence of identical circle segments, skirting along the wall. This type of orbit is called *skipping orbit*. The centre of the circularly shaped orbit travels with a (average) velocity v_d determined by the Fermi velocity and the distance of the centre from the wall. In particular orbit T_2 has a velocity component that *increases* the mean velocity of the electron, thus leading to an increase of the transported current *along this wall*. We will return to these peculiar edge trajectories in even greater detail when discussing quantum ballistic effects in chapter 4. The sequence of "hits and segments" continues until a bulk scattering event breaks the translation symmetry, and depending on the details of the event, either a new skipping

orbit or a bulk orbit (of the type T_1) will result. It is important to note that, whatever the particular shape of a skipping orbit along edge W_1 may be, one *always* can define a *complementary* skipping orbit at opposite edge W_2, and in figure 2.4 the complement to trajectory T_2 is indicated by T_3. Note that the mean velocity of the centre of orbit T_3 at wall W_2 is *opposite* to that of a trajectory at wall W_1!
Now what is the effect of these complementary trajectories on the *total or net* transport in the x-direction along the channel? From the symmetry of the system it is straightforward to see that the *local increase* of the current via orbits along wall W_1 is *exactly cancelled* by the *local decrease* along wall W_2.
In this discussion we assumed that the bulk did not contain (elastic) scatterers. It is not difficult to see that adding such scatterers does not affect the line of reasoning. So the cancellation of the two contributions at the edges continues to hold, irrespective of the addition of scatterers.
From this discussion two important conclusions can be drawn for this case of specular edge reflection. First, the two facing walls defining the channel each contribute an *equal but opposite additional contribution to the current.* This implies that no effect results for *the net total current* that flows through the channel, i.e. *the magneto transport of a narrow system is not affected by specular reflection at the walls.* However we may proceed one step further. The locally enhanced current at one wall which is compensated by an exactly as large reduced current at the facing wall, can be described as if, in addition to the *net current I* flowing through the sample, a *circulating current* is set up along the *whole periphery* of the sample. Note that this peculiar current certainly is affected by scattering events but it is not suppressed to zero!
It will not be difficult to see that this circulating current will continue to flow even if the net current is put to zero. So, the circulating edge current is only controlled by the strength of the magnetic field (and, to some extent by scattering), and it is *stationary and time-independent*, implying it to be *dissipation free*. The *persistent* nature of the current makes it a *thermodynamic equilibrium property* of the system as a whole. This persistent current has various similarities to the supercurrent flowing in a superconductor. We return to persistent currents in mesoscopic normal (i.e., non-superconducting) conductors in chapter 5.

Problem: in order for a persistent current of the type discussed above to develop, are there conditions for the inelastic scattering length that have to be fulfilled?

b. Diffusive wall reflection

The effect of the elastic scattering at the walls of the (2D) sample is depicted in figure 2.5. To demonstrate the effects most clearly we assume that the width W of the sample is comparable to the elastic scattering length l_e; other (limiting) cases can be deduced straightforwardly. Figure 2.5 shows three characteristic paths, taken at different values of the magnetic field B. Trajectory T_1, taken at zero field, demonstrates the effective decrease of the elastic mean free path as discussed before (see I.b), quantified by eq. (2.15). Increasing the field will reduce the cyclotron radius R_c, following eq. (1.8c). To demonstrate its consequences for electron transport two characteristic cases are shown.

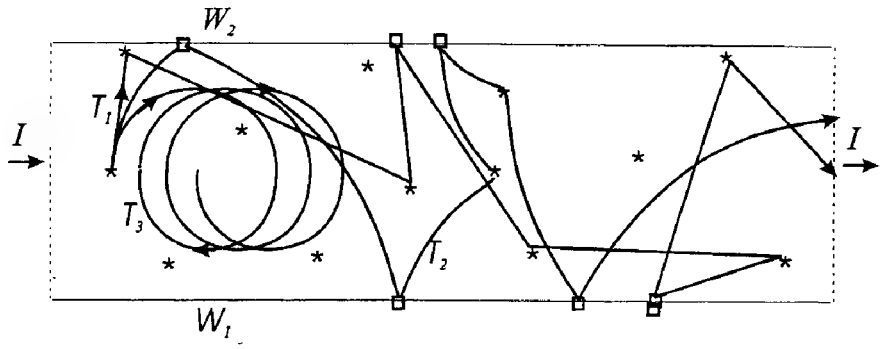


*Figure 2.5. Electrons moving in a magnetic field and between two diffusively scattering walls W_1 and W_2; * indicate scatterers. For clarity the channel width W is taken comparable to the elastic scattering length. The magnetic field is taken perpendicular to the plane of the picture. Trajectory T_1 is at zero magnetic field. The other two trajectories are for B-field values such that the associated cyclotron radii meet the conditions R_c~W (T_2) and R_c<<W (T_3) respectively..*

T_2 results for a field B_2 such that $R_c(B_2)$~W while T_3 results for the case that the field B_3 is considerably larger, i.e. $R_c(B_2)$<<W. At B_2 there is an increased probability for the electron to hit the walls. The increased elastic scattering at the walls yields a reduction of the effective mean free path and so the conductance will be reduced as well. Increasing the field to B_3 reduces the cyclotron diameter to well below the channel width. This will imply that the majority of the orbits will not experience the walls any more while traversing the length of the sample. This implies that for such (large) fields the character of the reflection at the walls becomes unimportant. We thus no longer can discriminate between specular and diffusive scattering and just the bulk properties of the channel determine the conductance.

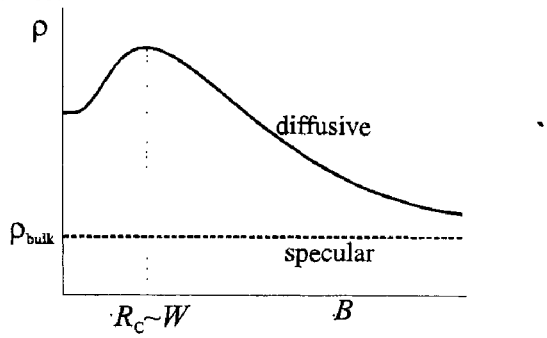


Figure 2.6. Classical magnetic field dependence of the longitudinal resistance of a channel demonstrating the effect of the nature of the wall scattering. The top curve shows the behaviour for diffusive scattering at the walls, with the characteristic bump occurring at a B-field such that the cyclotron radius matches the width of the channel; the bottom curve shows the flat, field-independent behaviour in case of fully specular reflection at the walls.

This is shown in figure 2.6. The topmost curve sketches the B-dependence of the resistance for diffusive wall scattering, clearly showing the initial rise of the resistance due to orbits of the type T_2 (see figure 2.5) with its peak at a field $B_c \sim B_2$. More accurate calculations show that the peak actually occurs at such magnetic field that $W=0.5R_c$ (see ref. 1 at the end of this chapter). This is followed by a decrease due to the reduction of the fraction of orbits experiencing the walls, with the resistance approaching the bulk value. Note that the resistance for ideal specular reflection (lower curve of figure 2.6) is independent of the field, as discussed in I.a and II.a.

Now we are in the position to see if the theoretical predictions made above are confirmed by experiments. Figure 2.7 shows results obtained by Thornton et.al. (Phys. Rev. Lett. **63**, 2128 (1989)) in a 2DEG in a GaAs/AlGaAs heterostructure (see chapter 1). Samples of different widths W between 0.13 and 1.15μm are investigated. To ensure diffusive scattering at the walls the structures are defined by ion implantation, a method which

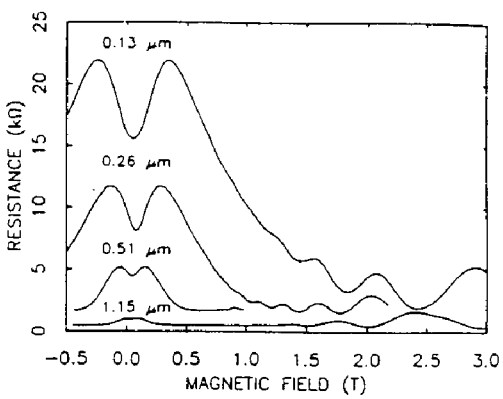


Figure 2.7. Experimental magnetic field dependence of the resistance of narrow channels, defined by ion beam implantation in a 2DEG in GaAs/AlGaAs heterostructure. The peculiar shape convincingly demonstrates the influence of diffusive wall scattering, as predicted and sketched in figure 2.6

deliberately induces local damage and depletion in the 2DEG, leading to rough walls. In the field range between 0 and 1T we very nicely find the characteristic behaviour sketched in figure 2.6, and a quantitative analyses shows also excellent agreement as far as the position of the peak of the resistance at $R_c(B_2) \approx W$.

Problem: check this agreement by taking the electron density from the paper of Thornton et.al.

Problem: imagine the channel would be made of a metal, with a Fermi energy of 1 eV. Estimate the magnetic field required to fulfil the condition for the peak in the resistance.

With this example we close this section on classical diffusive transport, including some classical size effects induced by confining the electrons into a narrow channel. In the next section we will see what effects may arise once we allow quantum interference to affect electron transport.

2.3 Quantum diffusive transport: weak localisation

In the preceding section 2.2 we discussed classical diffusive transport, i.e. in the limit where the phase coherence length l_φ (which we, as usual, loosely "mix" with the inelastic length l_i) is very much smaller than the size of the system. In this section we will discuss the opposite limit, defined by

$$\lambda_F \, , \, l_e << L \, , \, l_{i,\varphi}$$

In chapter 1 we discussed the fundamentally different nature of inelastic (or phase breaking) and elastic (non-phase breaking) scattering. The preservation of the phase as occurring in elastic scattering will maintain quantum interference, which will be shown to lead to marked effects in the conductance in this regime. Evidently to experimentally enter this regime we should bring the system at low temperatures to suppress phase-breaking influences like phonons.

The phenomenon we will concentrate on goes under the name of *weak localisation* (WL). Basically it results from the (quantum induced) *enhanced probability* for electrons experiencing many elastic scatterings to *return to their initial position*. This leads to a tendency for electrons to "stay where they are", i.e. some kind of electron localisation, resulting in a reduction of the conductance of the system.

2.3.a. *A simple introduction to the theory of quantum corrections to the conductance: Weak Localisation effect*

In deriving the Drude expression for the conductivity, eq. (2.3), a fundamental assumption is that each scattering event fully destroys all information on the initial velocity (size and direction) of the particle. In quantum terms this also implies that the phase correlation is completely suppressed during such event. If however, in contrast, the phase of the electron is (partially) preserved during scattering, this will affect the diffusion coefficient and so the conductance employing the Einstein relation eq. (2.8). From this equation, which contains the diffusion coefficient D given by the Kubo expression (2.10), it is immediately clear that conservation of the phase will introduce a correlation between the velocities for times *beyond the elastic scattering time* τ_e *and up to the phase-coherence time* τ_φ.

A fundamentally correct derivation of weak localisation requires a rather involved mathematical approach based on so-called diagrammatic techniques, which is well beyond the scope of our introductory text. We will use an approach employing "ray or geometric optics", augmented by the phase per raytrace, called the *Feynman path* method. Figure 2.8 shows the basic ingredients of this approach. The transport of an electron from an initial position 1 to a final position 2 is described as the transmission of an electron wave from 1 to 2 through the diffusive medium. This implies that the electron wave, starting at position 1, will become distributed over a number of *partial waves* ψ_j, each traversing the system along a different path p_j while experiencing many *elastic* scattering events before ultimately arriving at 2.

It should be noted that, to allow the use of the semi-classical Feynman paths to evaluate the total transmission probability, we implicitly assume that the electron really can follow a well-defined path, i.e. it should not be scattered too strongly on its characteristic length

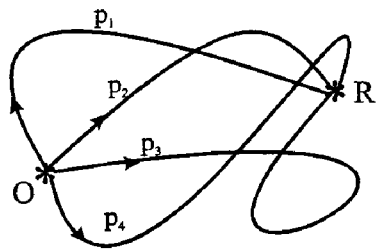


Figure 2.8. Electron waves travelling through a disordered medium containing elastic scatterers. The wave of a single electron at position O is distributed over a number of partial waves ψ_j, each following a well-defined path p_j, arriving at position R with a transmission coefficient t_j. Note that the phase φ_j on arriving at position R depends on the path length thus differing for each individual path.

scale, λ_F. As discussed immediately following eq. (2.5), violating this condition will lead to strong localisation, a phenomenon requiring a completely different description. To calculate the total quantum mechanical probability for the electron to travel from *O=>R* we have to evaluate the *vector sum* of all the partial waves arriving at *R*, i.e. including the *phases* φ_j of all the partial waves as acquired while following their individual paths. Figure 2.9 shows a scattering medium with a number of scatterers *, denoted by 0, 1, 2, 3, ...The electron wave is taken to start at position 0. We can distinguish between two different sets or ensembles of paths. The first set just connects two random positions, e.g. 0 and 7. Paths like {0,1,7}, {0,4,2,7}, {0,5,4,3,2,7} etc. all contribute to the total amplitude of the wave at 7, which then is given by

$$\Psi(7) = \sum_j \psi_j = \sum_j t_j \exp(i\varphi_j) \qquad (2.17)$$

Note that such a sum also contains contributions from multiple scattering, e.g. a path such as {0,4,2,4,2,7} is included, and so evaluating it will be a matter of tedious and accurate bookkeeping!

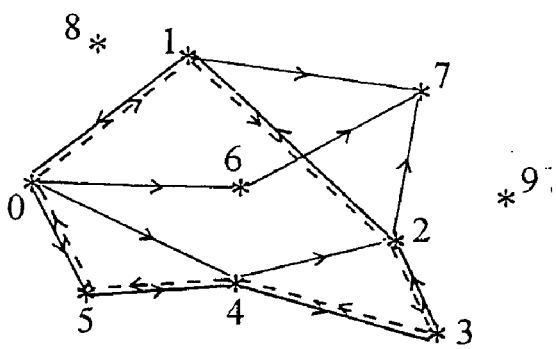


Figure 2.9. An electron wave starting at position 0, with its partial waves travelling along various paths. The solid lines show paths leading from 0 to (a randomly chosen final) position 7. The dashed sequence of lines forms a path that allows the electron to return to its starting position 0.

Formally, from eq. (2.17) we straightforwardly can write down the total probability for the wave at (say) position 7, as

$$P(7) \equiv |\Psi(7)|^2 = \sum_j t_j^2 + \sum_{j \neq k} t_j t_k \cos(\varphi_j - \varphi_k) \qquad (2.18)$$

Now we assume that *many different paths* contribute to the total amplitude. In the second term of eq. (2.18), containing the cross-product terms, the *random distribution* of the phases φ_j for all the paths p_i yields on average as many positive and negative contributions, thus reducing its sum to a small value. So, the total probability for forward transfer between two randomly chosen positions (0 and 7 in our case) approximately is just the sum of the probabilities of the individual partial waves.

$$P_{fwrd} \approx \sum_j t_j^2 \qquad (2.19)$$

Next we pose the question what happens if we take the initial and final position in space to be *the same*. This implies that we want to know what probability is associated with *returning* to the initial point. In figure 2.9 paths like e.g. {0,1,7,6,0} and {0,5,4,0} are typical examples, taken for the starting position 0. Let us look in a little more detail to the various trajectories involved in the total return probability. Concentrating first on randomly chosen *different* paths, it is evident that the same argument holds concerning the phases as given before for the different points in space and thus the total probability resulting from these trajectories is again the simple sum of the partial probabilities (eq. (2.19)).

However, in addition to these randomly chosen paths we also have to include a particular type of *complementary* trajectories. More specifically a typical example in figure 2.9 is {0,1,2,3,4,5,0} and {0,5,4,3,2,1,0}, i.e. the first is the *time-reversed* version of the second (and vice-versa). These time-reversed twins of course are not independent. In particular the identical length of the two paths implies that the phase-differences acquired during travel will be exactly the same. In addition the symmetry of elastic scattering events dictates that also the amplitude of the two partial waves will be the same.

Write j_{j+} for the amplitude of the wave along path p_j taken in one particular direction (say, clockwise), and t_{j-} for the arriving amplitude after traversing the path in the opposite, counter-clockwise direction. For the total contribution of this path to the return probability we thus find

$$P_j(0 \rightarrow 0) = t_{j+}^2 + t_{j-}^2 + t_{j+} t_{j-} \cos(\varphi_j - \varphi j + t_{j-} t_{j+} \cos(\varphi_j - \varphi_j) = 4t_j^2 \quad (2.20a)$$

as $t_{j+} = t_{j-} = t_j$ from the symmetry argument. For the results after summing the contributions of all paths we thus find

$$P_{return} \equiv P(0 \rightarrow 0) = \sum_j P_j(0 \rightarrow 0) = 4 \sum_j t_j^2 \qquad (2.20b)$$

One factor of 2 results from the fact that we simply have summed over two trajectories per entry j: for each path denoted by a single j also the complement was taken in the sum. However then we are still left with a second factor of 2. This factor is the direct consequence of the coherent, in-phase addition of time-reversed paths.

So we conclude that the *phase-coherent summation of time-reversed trajectories in a diffusive medium leads to an increased probability for electrons to return to their initial position.* This is commonly referred to as *coherent back scattering.* It implies that electrons indeed tend to remain at their initial site, an effect which leads to a reduction of the conductance (and an increase of the resistance) of the system, denoted as *Weak Localisation (WL).*

Before providing a more quantitative estimation of the strength of weak localisation in the conductance we first want to demonstrate a quite remarkable effect associated with the directions of the electron waves involved in this quantum coherent return phenomenon. Figure 2.10 shows two specific scatterers 1 and 2, with a plane electron wave incident on the two from the left. The wave, with the incoming plane wave front at A indicated by the dashed line, is (partially) scattered directly at each scatterer and returned, but in addition it is (partially) scattered to the other scatterer at a distance Δr_{12}, followed by returning to the left. To calculate the total leftwards outgoing wave we have to (coherently) sum the individual contributions, including the phases, and see what condition is required to recover a plane wave front at B and C, denoted by the dashed lines. Now consider the two parallel rays R1 (top) and R2 (bottom) of the incoming beam. R1 is scattered from 2 to 1 and we assume it to leave 1 at an angle δ relative to the incoming direction. The lower ray R2 follows the path 1 to 2 (i.e. reverse as compared to R1) and we look at the ray leaving 2 at the same angle δ. Note that the figure also shows the rays leaving at the complementary angle $-\delta$.
The difference in pathlength, $\Delta l_{12}(\delta)$, between the two rays that leave 2 and 1 immediately follows from the geometry, yielding

$$\Delta l_{12}(\delta) = a - a' = \Delta r_{12}(\cos(\theta) - \cos(\theta')) \qquad (2.21a)$$

with $\theta' = \theta + \delta$ (and $\theta'' - \delta$ for the complement) is the angle between Δr_{12} and the direction

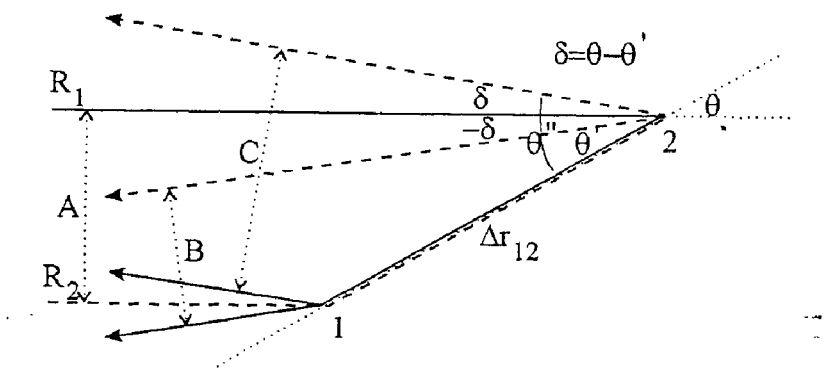


Figure 2.10. Coherent back-scattering of a plane electron wave at A incident on two elastic scatterers 1 and 2. At B a wave front leaving the scatterers at an angle $-\delta$ relative to the incoming beam is shown; similarly a wave front leaving at $+\delta$ is indicated at C.

of the incoming wave (see figure 2.10). Assuming δ to be small (see below) we find for the associated difference in phase

$$\Delta\varphi_{12}(\pm\delta) = 2\pi\frac{\Delta l_{12}(\pm\delta)}{\lambda_F} \approx 2\pi\frac{\Delta r_{12}}{\lambda_F}\left(-\frac{1}{2}\delta^2\cos(\theta) \pm \delta\sin(\theta)\right) \qquad (2.21b)$$

For the *total* outgoing wave we have to sum the + and - terms and so the contributions due to the sin(θ) term cancel, so we only have to consider the cosine term. The contributions of the two rays only add constructively if they differ in phase by no more than ~$\pi/2$. As the alignment of the two scatterers is completely random, we now have to average over all possible directions of Δr_{12} relative to the incoming beam, i.e. with the angle θ covering the range $-\pi/2<\theta<\pi/2$. This yields for the average over angle

$$<\Delta r_{12}\cos(\theta)> \approx \Delta r_{12}/2 \approx l_e/2 \qquad (2.22)$$

the last equality resulting from the fact that we take the two scatterers at their minimum average separation $\Delta r_{12} \sim l_e$. Combining eq. (2.22) with eq. (2.21b) and including the limit $<\Delta\varphi_{12}> \leq \pi/2$ we find the following maximum angle δ_{max} for the *reflected outgoing wave* relative to the incoming wave

$$\delta_{max} \approx \sqrt{\lambda_F / l_e} \qquad (2.23)$$

As discussed previously we limit ourselves to the case that the electron can travel freely at least over a number of wavelengths or $\lambda_F << l_e$, which implies that $\delta_{max} << 1$. So we have to conclude that the reflected wave only has an appreciable intensity within a (2D or 3D) cone with a (small) top-angle $2\delta_{max}$ relative to the direction of the incoming beam. This underlines that indeed the wave is truly *back-reflected* in a *coherent* way.

Now we are in the position to calculate quantitatively the effect of weak localisation on the conductance of the system. For the sake of simplicity we will limit the discussion to the 2D case. We employ the celebrated Einstein relation, eq. (2.8), for the conductivity and see how the coherent backreflection reduces the diffusion coefficient by an amount ΔD.
Using the Kubo expression (2.10) for D as the velocity-velocity correlation,

$$D = \int_0^\infty <\vec{v}_x(t)\vec{v}_x(0)> dt$$

we evaluate the correction resulting from the increased correlation due to the coherent backscattering, with $|v_x(t)| \sim |v_x(0)| \sim v_F$. The (negative) contribution to D thus can be written as

$$\Delta D \approx -<v_F v_F * \delta_{max}^2 * P_{return}(\Delta S(\vec{r} = \vec{0}))>_{direction} \qquad (2.24)$$

with $<....>_{direction}$ denoting the averaging over all (2D) directions of the incoming beam, as usual for the calculation of the diffusion coefficient. It leads to the $1/d$-factor, with d denoting the dimensionality of the problem (see also eq. (2.9)). The maximum angle δ_{max} is included squared as we are calculating the return *probability*. From our semi-classical ray-optics picture $P_{return}(\Delta S(r=0))$ is the (classical) probability for the particle, starting at

a certain position r=0, to return to within an area of size ΔS around this initial position, and within a time scale that is relevant for the problem at hand; note that this "origin" can be taken completely arbitrarily. This probability integrated over all times can thus be written

$$P_{return}(\Delta S(\vec{r}=\vec{0})) = \Delta S * C_{return} = \Delta S \int_0^\infty C(\vec{0},t)dt \equiv \Delta S \frac{1}{n_0}\int_0^\infty n(\vec{0},t)dt \qquad (2.25)$$

with $n(0,t)$ given by expression (2.12), i.e. representing the probability to return to the initial position at a certain time t. From (2.12) we immediate find for the 2D case

$$n_{2D}(\vec{0},t) \equiv n_0 C_{2D}(\vec{0},t) = \frac{n_0}{4\pi D t} \qquad (2.26)$$

Before proceeding by evaluating the integral of (2.25) we have to include two more terms to it, based on the relevant time scales in the problem. First, we have to realise that electrons traveling for times *shorter* than the elastic scattering time $\tau_e = l_e/v_F$ will not be able to return what so ever, and so they will not contribute to the backscattering. This might be accounted for by changing the lower limit to the integral from 0 to τ_e, but to allow for the statistical nature of this aspects it is covered by multiplying $C(0,t)$ by the factor $(1-\exp(-t/\tau_e))$. Secondly, there is also an upper bound to the integral: by the time the wave loses its phase coherence it no longer can contribute to *coherent* backscattering and so we should limit the integral to times smaller than the phase breaking time τ_φ. Again by accounting for the statistical nature it is incorporated by multiplying by the factor $\exp(-t/\tau_\varphi)$.

The last term to be defined in eq. (2.25) is the return area ΔS. From the discussion around the figures 2.9 and 2.10 it is clear that we can not define "a return" any better than given by the average distance between elastic scatterers $\Delta r_{12} \sim l_e$, and so we immediately find (in 2D) for the typical return area $\Delta S \sim l_e^2$.

Introducing the three factors into eq. (2.25) and including also eq. (2.26) we find

$$\Delta D_{2D} \approx -\frac{l_e^2 v_F^2 \delta_{max}}{2*4\pi D}\int_0^\infty \frac{1}{t} exp(-\frac{t}{\tau_\varphi})(1-exp(-\frac{t}{\tau_e}))dt \qquad (2.27)$$

The time-integral yields $\ln(1+\tau_\varphi/\tau_e)$, which equals $\sim\ln(\tau_\varphi/\tau_e)$ if the phase breaking time is much larger than the elastic time.

Thus the final expression for the weak localisation correction to the conductance reads

$$\Delta\sigma_{2D} = e^2 \rho_{2D} \Delta D_{2D} \approx -\frac{1}{4\pi}\frac{2e^2}{h} ln(\frac{\tau_\varphi}{\tau_e}) \qquad (2.28)$$

It demonstrates two important properties. First, its size is of the order of the conductance quantum e^2/h. Secondly it is only weakly dependent on the ratio of the phase breaking length and the elastic length, once this ratio is >>1.

Equation (2.28) is for the 2D case; similar expressions can be obtained for the 1D and 3D case. One finds that the effect strongly increases at a reduction of the dimensionality, i.e. it is largest in 1D.

It is also noteworthy to see that the dimensionality for weak localisation is governed by the phase breaking length l_φ in comparison to the size of the system.

Before proceeding to discuss some experimental results that demonstrate this quantum interference effect in the diffusive regime we first have to consider how one would be able to determine it: if the WL only reduces the conductivity by a small fraction we need a way to distinguish between the classical conductance (eq. (2.7a)) and the correction due to WL (eq. (2.28)). To this purpose we consider the effect of an applied magnetic field on the WL effect, and we will find that for sufficiently large fields the correction $\Delta_{WL}\sigma$ from eq. (2.28) becomes suppressed. This allows us to distinguish between the classical and the WL contribution.

Applying a magnetic field to an electron travelling at a velocity v_F results in two effects:

= Classically it will bend the path of the electron due to the Lorentz force, leading to the formation of the circular cyclotron orbits; we have discussed this in chapter 1, eqs. (1.8). The important question we have to pose is how this curvature affects the WL correction, i.e. what effect does it have on a trajectory and its time-reversed complement (see again figure 2.9). Evidently the field-induced curvature affects the precise shape of the trajectories. In particular, the two time-reversed trajectories will become increasingly dissimilar for increasing field strength. However, at the small fields required to affect the WL (see below) the effect due to the curvature turns out to be negligible.

= Quantum mechanically the magnetic field affects the canonical momentum p via the vector potential A defined by the magnetic field as $\vec{B} = \vec{\nabla} \wedge \vec{A}$. From introductory quantum mechanics we know that for a charge q=-e

$$\vec{p} = \hbar\vec{k} = m\vec{v} + q\vec{A} = m\vec{v} - e\vec{A} \tag{2.29}$$

which implies that the wavevector k becomes B-dependent. The electron travelling at the Fermi energy from position 1 to 2 along a path defined by $\vec{l}$ acquires a phase

$$\Delta\varphi_{1\to 2} = \int_1^2 \vec{k}\cdot d\vec{l} = 2\pi\frac{m}{h}\int_1^2 \vec{v}_F \cdot d\vec{l} - 2\pi\frac{e}{h}\int_1^2 \vec{A}\cdot d\vec{l} = \Delta_v\varphi_{1\to 2} + \Delta_A\varphi_{1\to 2} \tag{2.30}$$

Note that the phase acquired following the trajectory is affected by the magnetic field via the second term of eq. (2.30), i.e. via the *vector potential*. We will return to this aspect in more detail in chapter 5.

Now let us concentrate on a typical WL trajectory, i.e. a time-reversed pair that (approximately) returns to its initial position after a number of elastic scatterings (figure 2.11). The resulting closed trajectory implies that in the expression (2.30) for the phase the integral should be interpreted as a loop integral enclosing the area S. For the phase this yields

$$\Delta_{1\to 2}\varphi = k_F L - 2\pi\frac{e}{h}\oint \vec{A}\cdot d\vec{l} = k_F L - 2\pi\frac{e}{h}\iint \vec{B}\cdot d\vec{S} = k_F L - 2\pi\frac{e}{h}\Phi \tag{2.31}$$

with Φ denoting the *flux B.S penetrating the area enclose by the closed trajectory*. The quantity h/e is called the (single charge) *flux quantum*, written as Φ_0=h/e. Now take the two time-reversed trajectories. Encirculating it clockwise will yield

$$\Delta_{clws}\varphi = k_F L - 2\pi\frac{\Phi}{\Phi_0} \tag{2.32a}$$

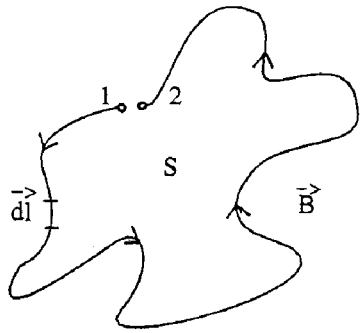


Figure 2.11. A typical trajectory that contributes to weak localisation, allowing an electron to leave from 1 and, after a number of elastic scattering events, returning to a position 2, or vice versa going from 2 to 1; 1 and 2 are separated by less than an elastic length l_e. The trajectory encloses an area indicated by S.

while traversing it in the opposite (i.e. counter clockwise) direction yields

$$\Delta_{cclws}\varphi = k_F L + 2\pi\frac{\Phi}{\Phi_0} \qquad (2.32b)$$

Note that the zero-field term $k_F L$ is (evidently) independent of the direction of travel. For the *difference* in phases acquired in travelling clockwise and counter clockwise along the time-reversed paths we thus find

$$\Delta\varphi = \Delta\varphi(B) = (2\pi - (-2\pi))\frac{\Phi}{\Phi_0} = 4\pi\frac{\Phi}{\Phi_0} = 4\pi\frac{BS}{\Phi_0} \qquad (2.33)$$

Note that the magnetic field enforces the two time-reversed trajectories to become dissimilar as far as their phases are concerned. Stated differently, a magnetic field leads to *time-reversal symmetry breaking*, a very general quantum phenomenon associated with the application of magnetic fields.

As we have discussed in relation to eq. (2.20) at zero B-field the fully *in-phase* coherent summation of the waves of the two time-reversed paths resulted in the factor-of-2 increase of the probability for back scattering. This implies that the phase-shift (2.33) induced by the field will *reduce* this sum following $\cos(\Delta\varphi(B))$, *i.e. the magnetic field indeed affects the contribution to the weak localisation.* In particular, at such a field that this shift amounts to $\sim\pi/2$ the (zero field) constructively interfering contribution to WL (from that particular path) at a certain B-field will turn into a destructively interfering contribution, and so it will become suppressed.
Now we have to realise that the total WL effect results from contributions from a large ensemble of individual (pairs of time-reversed) trajectories, each of different shape and size and so enclosed areas S_j. At a given field the phase shift depends (linearly) on the area S_j associated with each particular trajectory p_j. More specifically, for the whole set of trajectories the longer ones, which most likely (but not necessarily!) will also enclose the larger areas S, will acquire a larger phase shift than the shorter ones. The longest paths which can contribute to WL are those with a size of approximately the phase coherence

length, i.e. $L_{max}<\sim l_\varphi$, with an associated enclosed area $S_{l_\varphi} \approx \pi(l_\varphi / 2\pi)^2$. These will yield the largest phase shift at a given field. So, as a consequence, and based on eq. (2.33), weak localisation will start to become suppressed whenever

$$\Delta\varphi \approx \frac{\pi}{2} \approx \frac{2\pi}{\Phi_0} S_{l_\varphi} B_c = \frac{2\pi}{\Phi_0} \pi (\frac{l_\varphi}{2\pi})^2 B_c \qquad (2.34a)$$

From this a characteristic field B_c can be defined where the suppression starts to become effective

$$B_c \sim \frac{\Phi_0}{l_\varphi^2} = \frac{h}{e l_\varphi^2} \qquad (2.34b)$$

So, in conclusion, *weak localisation can be suppressed by applying a magnetic field such that the zero-field time-reversed trajectories become dissimilar in acquired phase due to time-reversal symmetry breaking*. This suppression of WL provides a way to study the effect experimentally.

2.3.b. *Experiments on Weak Localisation*

In the following part we discus two experiments showing weak localisation in the conductance, one on a semiconductor system and one on a metal. In order to demonstrate the very general nature of coherent backscattering in disordered system we show a third experiment which is not associated with electrons but employs optical waves.

1. Weak localisation in two-dimensional electron gas in a Si-MOSFET.
We first concentrate on the effect in a 2-dimensional electron gas in a semiconductor structure. (D.J. Bishop et.al., Phys. Rev. B26, 773-779 (1982)). The 2DEG is formed at the interface of a Si-MOSFET, the gate of which allows the electron density to be controlled; in this way the diffusion constant can be varied via the elastic mean free path and the Fermi velocity (see eq. 2.9). Figure 2.11 shows one of the results obtained in the experiment (noisy curve, "overlapped" by the solid curve b; the (fitting) curves a, b and c result from a full theoretical treatment). It is obtained at T=100 mK and an electron density n_e~4.5*10^{16} m^{-2}. Note the characteristic field scale for reaching a "field-independent" value B_c~ a few times 0.1T (note: 1 kG=0.1T).

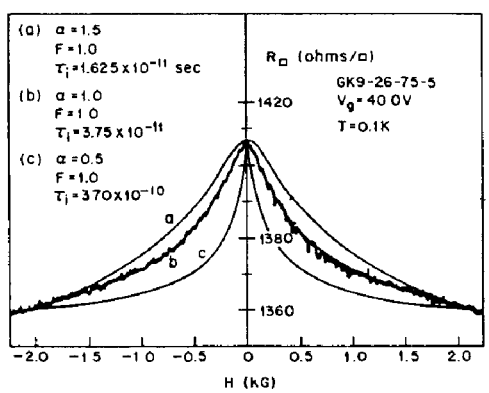


Figure 2.12. Reduction of the resistance with magnetic field, due to suppression of the weak localisation, in a 2D electron system at the Si-SiOx interface. The "noisy curve" is the actual experimental data. The smooth curves a, b and c represent different curves resulting from a fit to the theory.

Problem: Rewrite the resistance contribution due to WL in figure 2.12 in a change in the conductance; is the value thus found in reasonable agreement with eq. (2.28) assuming the logarithmic term to be of order unity?
Problem: Take from the original paper of Bishop et.al. the mobility. Evaluate the elastic mean free path from the electron density and the mobility. Evaluate also the phase coherence length from the characteristic field scale of fig. 2.12, using eq. (2.34b). Is the assumption of the near-unity value of the ln-term allowed?

2. Weak localisation in a thin metallic Mg film.
In this magnetoresistance experiment Bergmann (Phys. Rev. B25, 2937-2939 (1981); see also Physics Reports 107, 1-58 (1984): the figures are taken from this last paper, i.e. their figures 2.11. and 2.12) investigates the magnetic field dependence of the conductance of a very thin Mg film evaporated onto a crystalline quartz substrate. Homogeneous films with thicknesses varying between 7 and 12 nm were obtained by performing the evaporation with the substrate held at ~ 5K. The data points, shown in figure 2.13a, are taken at five different temperatures between 4.5 and 20K. The magnetic field for each temperature is indicated next to each individual experimental curve; it varies for the range of temperatures taken in this experiment. This change of characteristic field scale is

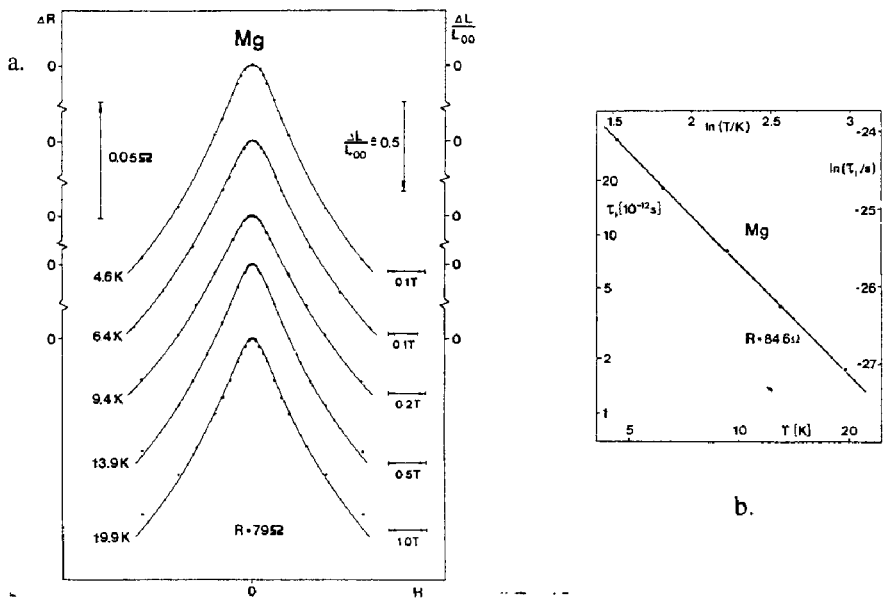


Figure 2.13. Weak localisation in a thin Mg film. Panel a. shows the experimental results obtained at five different temperatures between 4.6 and 19.9 K. The resistance in an applied magnetic field shows the characteristic behaviour of weak localisation, with a (temperature dependent) suppression field. In panel b. the phase breaking time is shown as deduced from the measurements.

accordance with the decrease of the phase breaking length with increasing temperature (eq. 2.34b). The full curves are fits to a full theoretical treatment of the WL problem; note the very good agreement obtained in this fit! From these results it is rather straightforward to evaluate the temperature dependence of the phase breaking time; this is shown in figure 2.13b.

Problem: Determine the dimensionality of this WL experiment. For this to do, remind yourself on which length scales are relevant. Take the needed values from the Phys. Rev paper of Bergmann.

3. Coherent backscattering in optics: the narrow coherent cone.-
The third experiment we want to discus seems to be of a completely different nature. In the preceding discussion we have concentrated on *electron* waves in diffusive systems. However, from the whole line of reasoning it is completely self-evident that the whole line of reasoning that leads to WL should hold for *any type of waves* in a diffusive medium! So it is anticipated that also mechanical waves (e.g. in acoustics) or electromagnetic waves should show the same phenomenon, with the (important!) exception that a magnetic field will not be able to break the time-reversal symmetry in these cases. This missing control mechanism (or "experimental knob") thus demands a different approach in order to investigate the phenomenon, in order to separate the WL contribution from the much larger "classical" background. This can be done by employing the aspect of coherent backscattering into a narrow cone as discussed before (see figure 2.10). We found the important result that a beam incident onto the diffusive medium shows an increase probability of being backreflected, preferentially within a cone

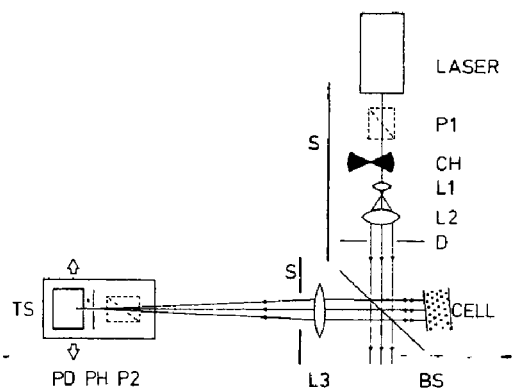


Figure 2.14. Schematic diagram of the optical set-up to study weak localisation of optical waves in a disordered medium.

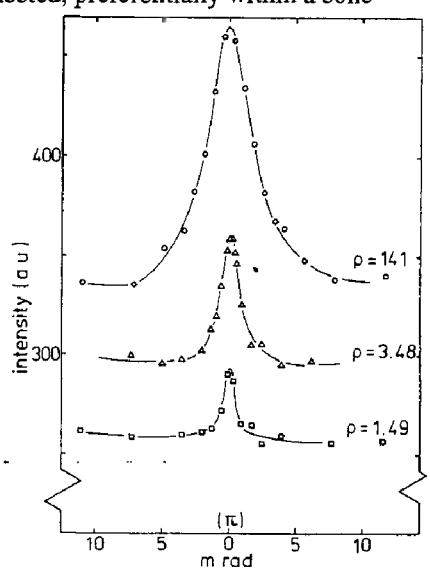


Figure 2.15. Coherent backscattering in a suspension containing 1 μm Latex spheres. ρ is the suspension density.

determined by eq. (2.23) with a total angle

$$\delta_{cone} = 2\delta_{max} \sim \sqrt{\lambda_F / l_e} \qquad (2.35)$$

Note that this expression is formally only "correct" for the 2D case; however it is found to depend only very weakly on dimensionality. We want to describe an optical (=electromagnetic wave) experiment which nicely demonstrates this coherent backscattering.

Figure 2.14 shows the set-up for this experiment (Albeda et.al., Phys. Rev. Lett. 55, 2692-2695 (1985)). Then optical wave is provided by a He-Ne laser. The diffusive medium is realised by dispersing Latex polystyrene spheres of ~ 1 μm diameter in water. The effective elastic mean free path could be varied by changing the the density of the suspension. The laser beam reaches the diffusive cell (which contains the "milky" suspension) via a beam splitter. The beam reflected from the cell is measured by a photo detector PD. The position of the detector can be varied in space, which allows an angular dependent determination of the reflected beam. The results of these measurements are shown in figure 2.14. The three sets of data are for different concentration of the particles in the suspension. The solid curves are fits to a more complete theory, but the typical angle is found to be in very good agreement with our simple eq. (2.23).

2.4 Closing remarks

In this chapter we have discussed a number of classical and one quantum phenomenon that are characteristic for mesoscopic systems in the diffusive regime. Here we want to make one additional remark derived from the quantum phenomenon of weak localisation. As we have shown this results from the complex electron wave interference in the diffusive background. It is of interest to briefly reconsider the magnitude of the effect, as given by eq. (2.28). Apart from the weak (logarithmic) dependence on the ratio of the times for phase breaking and elastic scattering, the typical amplitude in the change of the conductance due to WL amounts to e^2/h, i.e. the conductance quantum. Stated differently, irrespective of the *details* of the distribution of elastic scatterers (i.e., the *microscopic configuration*), the quantum effect on the conductance always is of the *same order of magnitude*. On the other hand, it will be clear that changing the configuration of the scatterers will result in some change in the conductance. Combining these arguments points towards the presumption that *varying the microscopic details of a mesoscopic diffusive medium while maintaining its macroscopic, average properties, leads to variations in the conductance that will be (typically) no larger than the conductance quantum.*

This conclusion can be verified in a much more rigid way, and the important and highly fundamental phenomenon of such fluctuations in the conductance of diffusive systems is very well known in the literature under the name of *Universal Conductance Fluctuations* or *UCF*. The adjective "universal" stresses the role of the conductance quantum in this respect. In addition to Weak Localisation it is the second important quantum phenomenon found in diffusive mesoscopic systems.

General references to the subject

1. "Quantum Transport in Semiconductor Nanostructures" by C.W.J. Beenakker and H. van Houten, in Solid State Physics, ed. H. Ehrenreich & D. Turnbull, Vol 44 (Academic Press Inc., 1991), sections 4, 5, 6 and 7.

3. Classical ballistic transport

Keywords: Electron reservoirs, transmission probability, Landauer approach, electron focussing, non-locality, negative Hall resistance

3.1 Introduction

In this chapter we want to address a less commonly experienced way of electron transport in solids. In chapter 2 we discussed the well-known diffusive regime: here the intense scattering of the electrons only allows us to calculate voltages and currents by evaluating the *average* motion of these particles, and not the detailed behaviour of a single electron. This also implied that (at least for the classical diffusive case) the currents are determined by the local electric fields inside the solid.
In contrast the regime we are concerned with in the present chapter is of a completely different nature. We will see that the total currents now can be obtained by concentrating on a single particle, whose time- and postion dependence we are able to describe accurately, and subsequently sum all particles involved.
Refering to section 3 of the first chapter the classical ballistic regime is characterised by the following relation between length scales:

$$\lambda_F << L < l_i, l_e$$

i.c. the size of the system, L, is taken to be considerably smaller than any of the characteristic scattering lengths. The Fermi wavelength is assumed to play no role at all: effects due to quantum interference will be discussed in the next chapter, 4. The implication of the relation of system size to scattering lengths implies that the electron, once set into motion, will continue in straight lines all the way through the whole system. It is this very simple way of motion, which is completely similar to the way larger (so by definition classically behaving) bodies like tennisballs, rockets and bullets move through free space, that defines the name of this regime as ballistic.
In this chapter we will see how this simple but less common propagation of electrons gives rise to rather surprising phenomena in transport properties like the resistance and the Hall voltage.
Section 3.2 will provide the basic ingredients to describe this classical ballistic regime, including a short introduction to the role of the probability of electron transmission and the associated Landauer approach in electron transport. Section 3.3 shows a few characteristic exemples of experimental results obtained in this area.

3.2 Transport mechanisms and description of the ballistic regime; the Landauer formalism

Figure 3.1 shows a system which demonstrates the most characteristic aspects of ballistic transport. A piece of conducting material of length L and width(s) W is assumed to be so clean that it does not contain any impurities etc. that may act as elastic scatterers for the electrons. In addition we assume that also no inelastic scattering takes place in the channel area, e.g. by taking the temperature sufficiently low. This immediately points towards a very important aspect of

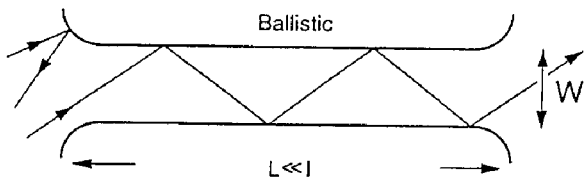


Figure 3.1. Particle moving through a narrow channel. As all scattering lengths are taken much larger than the length L and width W of the channel the electron trajectories are fully determined by the initial conditions (at the entrance) and the walls of the channel itself.

ballistic transport, namely the decoupling of the area that determines the resistance and the area of energy dissipation.
As figure 3.1 shows, electrons are allowed to traverse the system of interest without scattering. So, a particle entering the system from the left with a certain velocity, or kinetic energy, will maintain this while travelling all the way to the right. This implies that nowhere in the system of interest, as defined by the "boundaries or limits" L and W, the particle is allowed to exchange energy with the environment. However, it is evident that the particle has been set into motion, e.g. at some position to the left from the left entrance of the channel, by some type of source, e.g. a battery. As a "mirror image" of this process the electron will be absorbed by a "sink", finalising its independent ballistic life in a sea of brothers and/or sisters. It is at this place that the energy of the electron can be dissipated, via inelastic scattering processes like electron-electron or electron-phonon interaction. So we conclude that, *while the transport occurs inside our system of interest, the dissipation of energy takes place outside this restricted area.* This implies that for the description of ballistic transport processes and phenomena we are allowed to completely decouple the (interesting) transport system from the (dull and/or complicated) dissipation systems. This separation provides a tremendous simplification!

A second, related, aspect which surfaces in this discussion is that processes (like current flow) occurring at a certain position in space are strongly determined by the processes (e.g. injecting of particles by a battery) taking place at other, remote positions. This is called non-locality, and so ballistic processes are nearly always non-local in nature.

The separation of the transport area from the dissipation areas allows us to discuss the two independently. So, before proceeding with the transport, we will first concentrate on the dissipation areas.
From introductory solid state physics (and from chapter 1) we know that electrons behave as Fermions, obeying Fermi-Dirac (FD) statisties. They are allowed to move (more or less) freely throughout a piece of conducting material, with the velocity of "each individual" electron determined by the particular state it occupies. At zero temperature the maximum or Fermi velocity v_F, and likewise the maximum kinetic energy or Fermi energy E_F, is determined by the total electron density (see eqs. (1.14) and (1.16)). Likewise at non-zero temperature, the distribution of kinetic energies over the particles in a certain region in the solid is completely determined by the particle statistics (eq. (1.6)).

In thermal equilibrium each electron interacts with the environment, i.e. the phonons of the background atomic lattice as well as with the other electrons. From thermodynamics and statistical mechanics we know that if particles, contained in a certain volume, are in full equilibrium (i.e. there is complete exchange of particles throughout the whole volume) the ensemble of particles is characterised by a constant *chemical potential* μ_{chem}. It can be shown that for particles subject to FD-statistics the chemical potential equals the Fermi energy (at zero temperature), i.e. the maximum kinetic energy of the particles.
Such a region in space where electrons are in full mutual thermal equilibrium, characterised by a uniform chemical potential, is denoted an *electron reservoir* or *electron bath*. So, all particles inside a reservoir will have a kinetic energy $E_{\text{kin}}<\mu_{\text{res}}$, the chemical potential of the reservoir.

Problem: Inside a reservoir electrons are in full thermal equilibrium. What restrictions on the size of such a reservoir can be defined, in terms of the characteristic length scales from chapter 1?

In addition to the kinetic energy the particles may have potential energies as well. E.g., electrons in a reservoir kept at a certain *electrostatic potential* V, will maintain there kinetic energy but are able to do additional work by draining away power from the electrostatic source. So, the properties of the reservoir connected to a circuit will be completely determined by its *electro-chemical potential* μ given by

$$\mu = \mu_{chem} + qV = \mu_{chem} - eV \cong E_F - eV \qquad (3.1)$$

with the last near-equality being exact at zero temperature.

The way we have defined the electron reservoirs is quite restrictive: we assume all particles to be in mutual equilibrium. Evidently this will hold for particles inside a closed box, as these "meet one another" sufficiently to establish this overall state. It is, however, our aim to use these reservoirs as the sources of electrons in our ballistic systems, which will require that a reservoirs should somehow become connected to the external world, e.g. our ballistic channel system of figure 3.1. This immediately rises the question what will happen if we allow particles from outside to enter the box and, by the same token, particles from inside to leave the box.
Now we make the important assumption that all phenomena we want to study occur under *near-*

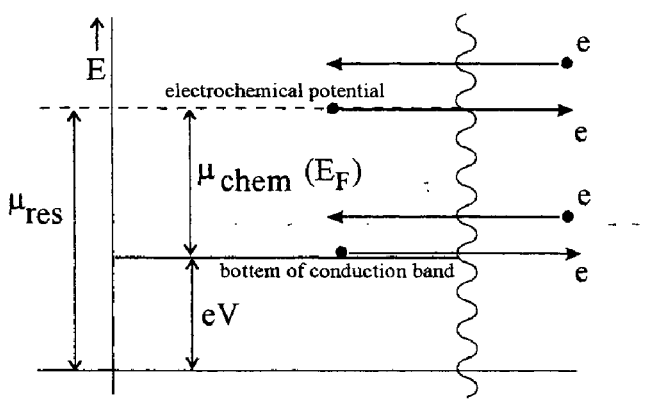
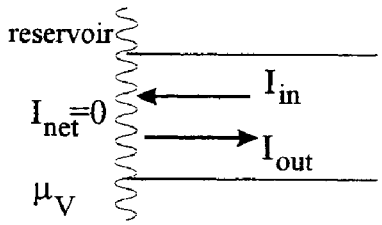


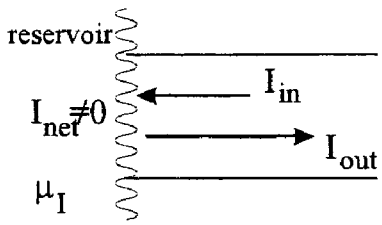
Figure 3.2. The energy diagramme of an electron reservoir. Note the chemical potential μ_{chem} ($\sim E_F$), and the total maximum energy given by the electro-chemical energy characterising the reservoir, μ_{res}. Electrons leave with energies $E<\mu_{res}$ while they can enter at any energy. The reservoir is bounded by the wavy line.



a.

Figure 3.4. Contacts and particle exchange. The outgoing current I_{out} is determined by the electrochemical potential μ_{res} of the reservoir. The incoming current I_{in} is determined by processes occurring inside the sample. The net current is the difference between the ingoing and outgoing currents.

a. Voltage contact: the average net current flowing through the contact is zero. The contact establishes this condition by appropriately adjusting its electrochmical potential μ_V.



b.

b. Current contact: the average net current flowing through the current contact is determined by the external current supply source. The electrochemical potential of the contact μ_I adjusts itself such as to make the average difference between the outgoing and incoming current equal to the externally applied net current I.

potential, in this way adjusting I_{out}. So, the condition at the voltage contact of figure 3.4a is

$$\mu_V \text{ such that } I_{net}=I_{out} - I_{in}=0 \text{ or } I_{out}=I_{in} \qquad (3.2a)$$

A similar reasoning is used for the current contact (figure 3.4b). As the total net current is established by the external current source, i.e. $I=I_{net}$, we immediately are lead to the conclusion that the electrochemical potential of the current contact μ_I adjusts itself to match the average current balance, i.e.

$$\mu_I \text{ such that } I_{net}=I_{out} - I_{in}=I \qquad (3.2b)$$

Note that the approach taken here is rather opposite to the way currents and voltages are calulated in classical diffusive systems of the type encountered in the macroscopic world. In that case we usually apply a finite voltage and evaluate the current resulting from it. In contrast in the case just discussed we impress the net-current condition and let the electrochemical potential adjust itself to provide this required current level.

From the discussion it is obvious that the common distinction made between current and voltage contacts no longer holds in the present case. Usually a voltage contact is said to couple to (and so affect) the sample only weakly, simply because of its "zero current measuring condition". As we

have seen this is not at all the case for the present voltage contact. The intensive exchange of particles between the contact and the sample does (virtually) not depend on the *net* current drawn through the contact and so voltage and current contacts of the same size equally interact with the sample. This implies that a description of the properties of the *sample* has to include the properties of the contacts, at least to the degree of how easily particles are exchanged with it, i.c. the degree of *transparency* or *transmittivity* of the sample-reservoir interface. Below we will see that the concept of particle transmission is of central importance in the description of electron transport in ballistic systems.

One question may arise after studying this section. In more common, i.c. diffusive, electronic systems (see chapter 2) the current(-density) at a certain position is determined by the (local) electric field. In contrast in the present discussion on ballistic systems it seems that the electric field does not play any role, neither local nor non-local. To a large extent this is a correct notion. In the diffusive case of chapter 2 the current density is given by the *average drift velocity* of the electrons. As the extensive scattering completely randomises the *direction of the instantanous total velocity* (which, for the electrons contributing to the conduction, approximately equals the Fermi velocity), the resulting drift velocity (which is much smaller than the Fermi velocity) only results from the acceleration in the local electric field during the short time interval between two successive scattering events.
This is not so for the ballistic case. To appreciate this we return to the discussion just following eq. (3.1). There the assumption was made that the currents in the system are such that the system is kept close to equilibrium. This for instance implies that the net current in a current contact is much smaller than its incoming and outgoing currents (see figure 3.4b); just as well it leads to the requirement that all contacts are only allowed to support electrochemical potentials much smaller than the Fermi energy. This implies that nowhere in the system the maximum kinetic energy of particles deviates appreciably from the local Fermi energy (or local chemical potential) (Note: it should be remarked here that in a ballistic system no single local Fermi energy can be defined, because of the anisotropy resulting from the external currents!!).
To be somewhat more specific assume the maximum difference in electrochemical potentials in the circuit is eV_{ext}, with $eV_{ext}<<E_F$. A particle leaving a contact at the Fermi energy and experiencing the electric fields associated with the electrochemical potential differences in the sample may show a maximum angular deviation form its initial direction (i.c. the direction obtained at the moment of injection from the contact) approximately given by $\delta_{max}\sim eV_{ext}/E_F<<1$. This means that a particle injected from a contact into the sample and moving at the Fermi velocity into a certain direction, will continue to do so all the way through the sample, without being affected appreceably by the externally applied electric field.
So in conclusion, while in diffusive systems electric fields determine the local (and overall) currentflow in the sample, this does not hold for ballistic systems.

The question which immediately arises is: if it is not the electric field which governs the particle flow, what then determines the currents in a ballistic system. To that purpose figure 3.5 shows a typical multi-terminal configuration, i.c. a sample with three contacts. Let us concentrate on the leftmost reservoir 1 and see how different particles, starting from this contact move through the sample, ultimately leaving the system via one of the contacts. Note the strong similarity to the well known "pin-ball" playing game. Assuming that the walls of the system show full specular

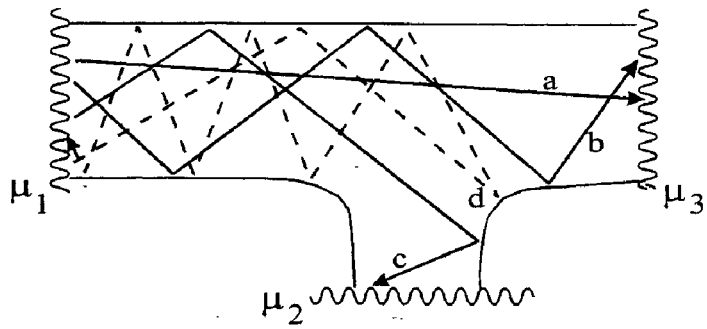


Figure 3.5. A typical 3-terminal ballistic structure to illustrate the role of the transmission probability. The central T-junction can be described by a scattering- or distributionmatrix for particles to and from all three reservoirs. Lines and arrows leaving the left reservoir 1 indicate four different particles a-d arriving at the different contacts, including one of these particles (d) being returned (= reflected) to contact 1.

reflection we can indicate the various paths a particle can follow after leaving contact 1. Particle *a* follows a path directly connecting contact 1 and 3. Path *b* is from a particle also arriving at contact 3, however only so after experiencing a number of (specular) reflections at the walls of the sample. Particle *c* is reflected a few times at the walls of the input lead before becoming scattered into the lead towards contact 2. Finally path *d* shows the case that the incoming particle experiences a number of reflections at different walls, ultimately becoming back-reflected to the contact of origin, 1. Based on these few examples of the "pin-ball" experiment we can conclude that the particle flow is determined by the *scattering properties* of the system. As the current flowing from contact k to contact m is determined by the number of particles leaving k and arriving at m, we can obtain the current flow by simply evaluating the *transmission probability* for particles arriving at m from k. This transmission probability for travelling from contact k to contact m, denoted by T_{mk}, is defined as the fraction of particles emitted from k that arrives at m; by definition $0<T_{mk}<1$. It can be determined for each pair of contacts by repeating the "pin-ball" experiment many times.

The use of transmission probabilities to calculate the currents at the terminals of a sample (and so the conductances between these terminals) is at the basis of the so-called *Landauer formalism*, named after R. Landauers original pioneering approach starting in 1957 (which, by the way, has been highly controversial for at least two decades, before becoming of major importance!). We will employ this method at many occasions throughout this book, both in the classical regime as well as for the quantum case. In this last case we feel considerably more at home, as in the quantum case the occurrence of a particle at some specific position in space is determined by the local amplitude of the wave function, a quantity which intrinsically represents its probability. In addition, to calculate such amplitudes one has to solve the Schroedinger wave-equation in space, which simply *is* a (wave-)scattering problem. We will return to this problem for the quantum mechanical case in Chapter 4 in the quantum ballistic regime, where we also will derive the specific relation between the conductance and the transmission probability, as well as how the currents are determined quantitatively by the electrochemical potentials at the different contacts of the sample system.

3.3 Experiments in the classical ballistic regime: electron focussing and negative-Hall resistance

The preceding section provides us with a very simple and transparent technique to understand and describe electron transport in samples in the ballistic regime. In the present section we want to discuss a few characteristic experiments in this regime, which show the strength of this ballistic scattering approach. In addition these experiments show results which are at first sight rather counterintuitive, in this way stressing the fundamentally different mechanisms being responsible for the electronic transport in the ballistic as compared to the (more common) diffusive regime.
We will discuss two types experiments: the first goes under the name of electron focussing and the second is referred to as the negative-Hall resistance.

a. Electron focussing

Figure 3.6 shows the sample lay-out for the electron focussing experiment. For simplicity we will limit ourselves to the case of a 2-dimensional electron gas (2DEG), as established in a semiconductor heterostructue, as discussed in Chapter 1 and in Appendix **X**. Remember that such 2DEG shows the required ballistic properties because of the use of remote doping. As seen in figure 3.6 the "infinite" 2D plane is divided into three areas (each with contacts/reservoirs attached to it) by some type of "line-screens", actually implemented by surface gates put at a negative potential realtive to the 2DEG (see figure 3.7 and Chapter 1 for more details). Small openings or point contacts are inserted in the screen between the lower left area and the top semi-infinite 2D

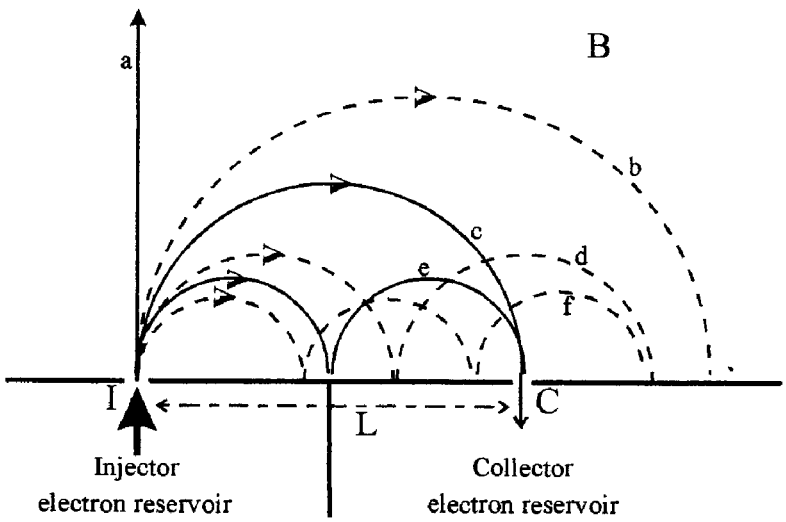


Figure 3.6. Electron focussing in a 2-dimensional electron gas. Electrons, only allowed to move in the 2D-plane of the figure, are injected from a narrow opening (or point contact) I in the "line-screen" from the injector reservoir into the semi-infinite area above the screen. With a magnetic field applied perpendicular to the 2D electron system, the injected electrons will follow the curved circular cyclotron orbits, with the diameter determined by the field strength B. At a second opening C in the screen, the collector, electrons can be detected and drained away via the collector reservoir.

plane, as well as between the lower right area and the semi-infinite plane. The two point contacts are separated by a distance L. We use the leftmost point contact I to inject electrons into the system (by applying a small (electrochemical) potential to the injector reservoir. By the same token the right point contact C is used to see if any electron arrives *at that particular position*. We can apply a magnetic field perpendicular to the 2DEG plane, i.e. always electrons will move perpendicular to the B-field.

For simplicity we assume that all electrons injected at I will do so perpendicularly to the screen (we will return to this condition later). At zero B-field the injected electron will set of into the semi-infinite 2DEG space along a straight line, indicated by a in figure 3.6 (we are in the ballistic regime!), and will ultimately be drained away by the (large) contact connected to that area. Once the magnetic field is switched on electrons will become bended away from their straight path and start to follow circular cyclotron orbits, as discussed in Chapter 1. From that discussion we know that the radius of curvature (or cyclotron radius) is given by eq. (1.6c) as $R_c=mv_F/eB$, i.e. it is inversely proportional to the magnetic field. Just as a typical value for a 2DEG in GaAs one finds R_c=1 μm at B=0.1 T. For the majority of chosen magnetic field values the circular orbit followed by the injected electrons will hit the screen at a position away from the collector opening c: see paths b, d and f. Only for very specific values the beam of electrons will be arrive at the collector, either directly (path c in figure 3.6) or indirectly after one or more (specular) reflection(s) at the screenwall (path e). It is straightforward to calculate the field values for this focussing to occur, as the condition is simply given by

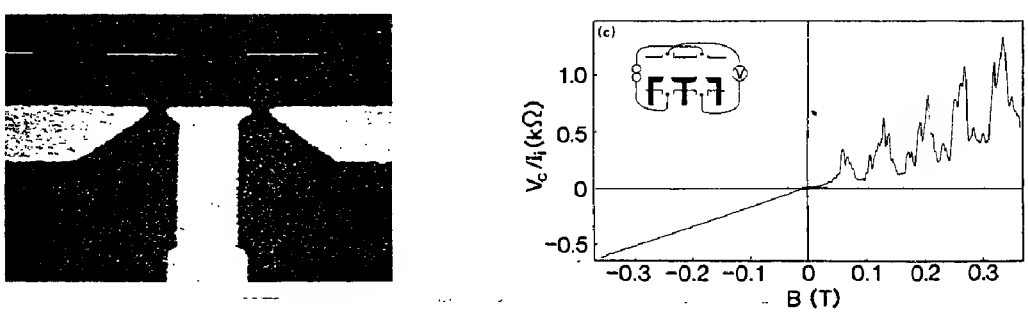
$$N*2R_c = L \qquad (N=1,2,3,\ldots) \tag{3.3a}$$

or

$$B_N = N*\frac{2mv_F}{eL} \tag{3.3b}$$

From this we see that the condition for focussing is fulfilled for magnetic field values at equal intervals $\Delta B=2mv_F/eL$.

Figure 3.7 shows the experiment. In figure 3.7a the layout of the actual sample is shown: the



a. b.

Figure 3.7. Experimental details and results for electron focussing in a 2DEG system. a shows the actual sample used for this experiment, with the two opening between the wedge-shaped gate ends forms the two point contacts for injection and detection. In b the voltage (or electrochemical potential) measured at contact C is displayed in dependence of the magnetic field; T=4K.

Scanning Electron Microscope (SEM) picture shows the surface gate pattern in lighter shade, with the two point contacts formed at the "meeting" of the sharp wedge-shaped features left and right from the central screen, which separates the injector and collector regions (top left and top right (or reversed, as this double point contact sample is completely symmetric)). Figure 3.7b displays the result obtained at T=4K (H. van Houten et.al., Phys. Rev. B34, 8556 (1989)). It exactly shows the behaviour predicted from the simple arguments formalised in eq. (3.3b), with peaks in the voltage measured at the detector contact, resulting from the focussing of the electron path onto the detector point contact C. Note the nice characteristic pattern of peaks at equal intervals in the field B. Note also the *absence* of the peaks at magnetic fields of opposite polarity.

This experiment only works out if the electrons can traverse the large 2D area without being scattered, i.e. truely ballistically. If any scattering occurs, either elastically or inelastically, this will change the direction of the travelling electron, in this way deviating it from the detector point contact. The main consequence of such scatterings will be to reduce the number of electrons arriving at the detector, and so the detected voltage will drop. So, if the 2DEG shows a finite mean free path l, we anticipate that the detector voltage will depend on it exponentially, approximately following $\exp(-\pi L/2l)$, i.e. determined by the ratio of the mean free path l to the travelled distance along the semi-circle(s) $\pi L/2$. Note that the travelled length is the same irrespective of the number of times N-1 the particle is reflected at the screenwall!

Problem: discuss the effect on the peak height due to a less-than-unity (specular) reflection at the screen; in particular consider this for different peaks, i.e. with different N.

The effect of scattering is shown in figure 3.8, an experiment performed on a multi-point contact

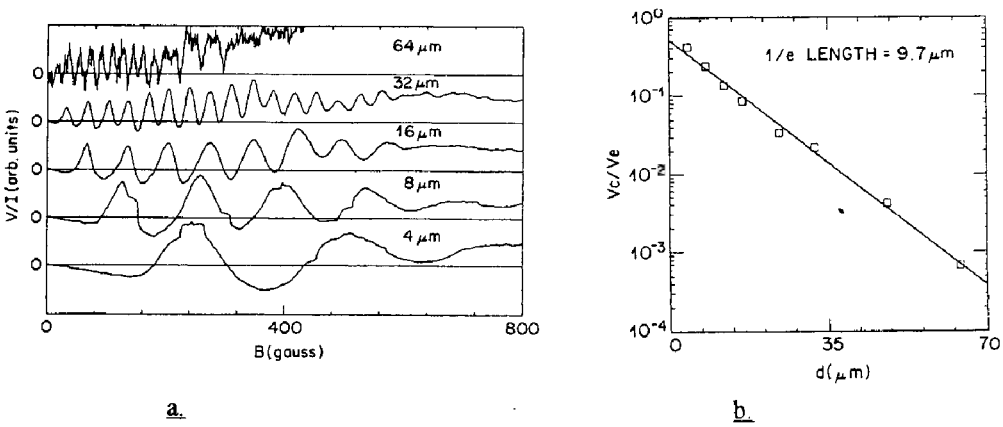


Figure 3.8. Electron focussing in a multi-point contact geometry, performed at T=0.3K. a shows the peak patterns obtained at the different detector point contacts are located at 4, 8, 16, 32 and 64 μm from the injector point contact. Note the reduction of the field interval between the peaks with increasing distance l, and the strong reduction of the amplitude of the peaks. b clearly demonstrates the exponential behaviour of the amplitude versus the distance.

geometry, with (detector) point contacts located at distances L=4 to 64(!) μm. The experiment, performed by J. Spector et.al. (Surf. Sc. **228**, 283 (1990)), clearly demonstrates the inverse proportionality of the intervals ΔB on the different lengths L as given by eq. (3.3b). In addition the very strong reduction of the amplitude is clearly visible. If the amplitude is presented (on a logarithmic scale) versus the distance between injector and detector, the exponential behaviour is striking, which confirms our intuitive guess presented before. From the exponent one finds a mean free paths $l \sim 15$ μm, which is approximately twice as small as obtained for the bulk of this 2DEG by measuring the mobility (see Chapter 1 for details). So, it seems that scattering is more effective in case of focussing as compared to normal transport. This is not too difficult to see, as for focussing the requirement is rather strict: any small deviation form the original path will make the electron to miss the detector opening, while in transport the electron contributes to the current als long as it keeps a forward velocity, even though the contribution will be reduced by somewhat like a cosine of the angle between the *average current* direction and the *actual velocity* direction.

Before closing the discussion of this experiment we have to return to the condition for perpendicular entrance of the electrons from the injector. Although we will not go into all details here it is obvious that electrons will (ballistically) traverse the point contact at various angles from the normal. From quantum mechanical arguments (see Chapter 4) we can show that indeed the electrons enters in a finite opening angle, the width of which is basically determined by interference effect occurring of the electrons with their Fermi-wavelength and the size (width) of the point contact. It is rather easy to see that this non-zero opening angle results in a broadening of the peaks. A finite angle simply implies that already at fields $B<B_N$ electrons entering at a non-zero angle (relative to perpendicular) can reach the detector and give rise to a current or voltage at contact C. More details can be found in the work of Beenakker and van Houten [1].

Problem: convince yourself from this argument by some simple schetches. In addition discuss what additional mechanisms may lead to a broadening of the peaks.

b. Negative Hall resistance

In this subsection we will discuss experiments which study transport in four-terminal geometries of a cross shape. Such configuration is particularly known for its use in Hall experiments, and one of the experiments we want to address relates to the question of how the ballistic nature of transport affects such a Hall experiment. The rather peculiar results provide a quite impressive way to show the differences between the two regimes, diffusive and ballistic.
For this purpose we assume a cross-shaped sample, as shown figure 3.9; again for reasons of simplicity we take a 2DEG. In figure 3.9a the sample is assumed to be *diffusive*, i.e. the size of the system (width and length) is much larger than the scattering length. Current is applied to the via current contacts 2 and 4, and to be specific, the last contact is put at zero (electrochemical) potential, i.e. $V_4=-\mu_4/e=0$ (note: $e=|e|$). If the current flows from contact 4 to contact 2, the electrons move oppositely from contact 2 to 4. We take the direction of the B-field relative to the diffusive electron flow such that the electrons are bent towards contact 3. If the cross is ideally symmetric, the resulting excess negative charge at 3 implies a *positive voltage difference* V_1-V_3,

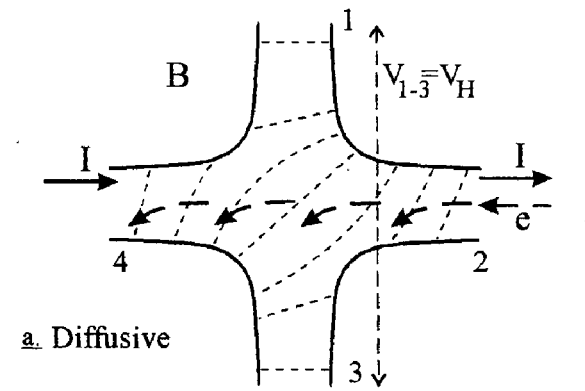


Figure 3.9a. The Hall effect of a 2D electron system in the diffusive regime. As usual two current contacts (2 and 4) are used for supplying the current; two voltage contacts (1 and 3) allow measurement of the Hall voltage via μ_1 - μ_3. Dashed lines are equipotential lines.

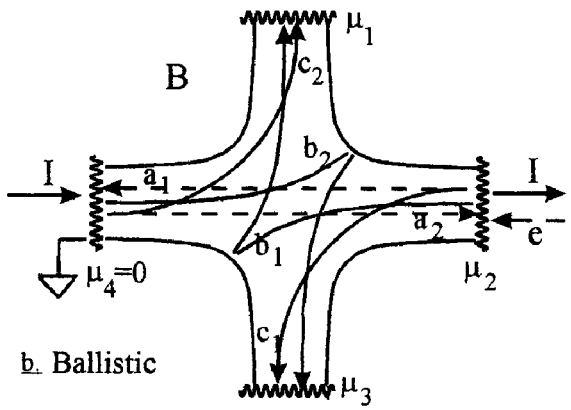


Figure 3.9b. The Hall effect in a 2DEG in the ballistic regime. The Hall voltage (from μ_1 and μ_3) now follows from the particular electron trajectories and the requirement of zero net current at the two voltage contacts 1 and 3. Three pairs of trajectories a,b and c are shown for increasing magnetic fields.

i.e.

$$R_{Hall,diff} \equiv \frac{V_{Hall,diff}}{I} \equiv \frac{V_1 - V_3}{I} = -\frac{\mu_1 - \mu_3}{|e|I} > 0 \qquad (3.4)$$

Now let us turn to the ballistic case, which we evaluate by looking at trajectories of electrons travelling at the Fermi velocity. To evaluate the resulting Hall potential in figure 3.9b three (pairs of) electron paths at different magnetic fields are shown, all emanating from the current contacts 2 and 4. Note that, following the same arguments as discussed when introducing the current- and voltage contacts in relation to figure 3.4, we *impress* the *net* current I flowing in and out the current contacts and we evaluate the electrochemical potentials that establish themseleves self-consistently at the various contacts.

The first pair of paths (*a1* from contact 2 and *a2* from contact 4) is taken at zero magnetic field: evidently the trajectories are straight lines, with *a1* being absorbed in reservoir 4 and *a2* in contact 2. Increasing the field may lead to trajectories like *b1* and *b2*. The peculiar property of these trajectories is that, despite being deflected "to the left", the specular reflection of the particle makes it to arrive at the contact terminal *at the opposite side of the sample*. This will make the charge to build up with opposite polarity as compared to the common diffusive case, thus yielding a *negative Hall voltage*. To be more specific for the case of these two trajectories, we employ the "zero net current" requirement for voltage contacts. For the case of simplicity we assume that for the given magnetic field *only* electrons that follow the indicated trajectory will reach contacts 1 and 3. First consider trajectory *b2*: as all electrons emitted from contact 4 have an energy $<\mu_4=0$ (the electrochemical potential of this lead) all electrons arriving at reservoir 3 will have this μ. As for voltage contact 3 the requirement is that just as many electrons leave the reservoir as there are absorbed, this immediately implies that the electrochemical potential of the reservoir will adjust itself to that of the incoming particle flow, i.c. $\mu_3=\mu_4=0$ and so $V_3=0$. Because of symmetry evidently the same argument holds for electrons on trajectory *b1*, leaving from reservior 2 and arriving at 1. This implies $\mu_1=\mu_2$. As current is forced from 4 to 2 this yields $V_2<V_4=0$ and so $\mu_2=-eV_2>0$. So we conclude that $\mu_1=\mu_2=-eV_2>0$ or $V_2<0$. This implies that the Hall voltage for the ballistic case and for this particular magnetic field value is given by

$$R_{Hall,ball} \equiv \frac{V_{Hall,ball}}{I} \equiv \frac{V_1 - V_3}{I} = -\frac{\mu_1 - \mu_3}{|e|I} \approx -\frac{\mu_2 - \mu_4}{|e|I} = \frac{V_2 - V_4}{I} < 0 \qquad (3.5)$$

So, for the ballistic case and at the given magnetic field we thus find a *negative Hall resistance*, opposite to the diffusive counterpart of figure 3.9a. Before switching to the actual experiment let us first see what happens if we increase the field further, i.c. trajectories *c1* and *c2*. As is obvious from figure 3.9b the electrons now arrive again at their "regular" contacts, and so the Hall voltage will be again positive. In summary, we anticipate in a small symmetric cross with specularly reflecting walls that for certain ranges in the magnetic field the Hall voltage will become negative.

Figure 3.10 shows the actual result of such an experiment (Ford at.al., Phys. Rev. Lett. **62**, 2724 (1989)) on a high-mobility 2DEG structure. In the field range between - and + 0.15T the Hall resistance is found to have the opposite sign as compared to the high field range.

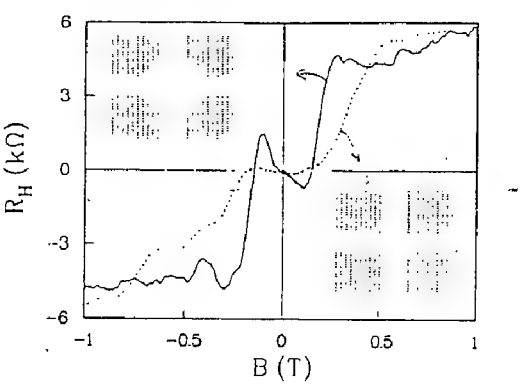


Figure 3.10. Experimental demonstration of the negative Hall resistance in a ballistc cross. Note the negative Hall range from ~ -0.15 to 0.15 T.

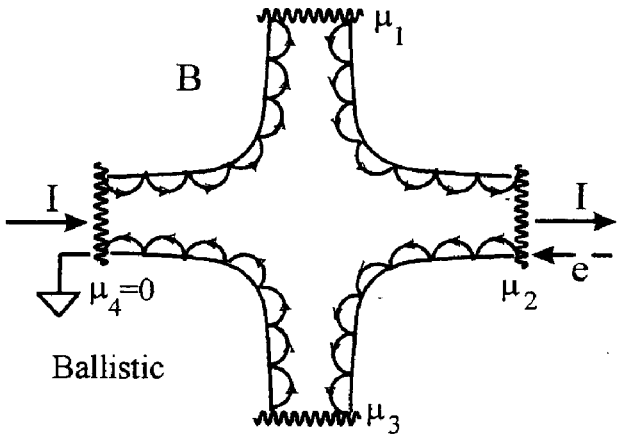


Figure 3.11. Hall effect in the ballistic regime at large magnetic fields. As the cyclotron radius is small compared to the system sizes the electrons contributing to the current will flow along the walls in "skipping orbits".

Before closing this chapter we want to pose the question on what will occur at such large magnetic fields that the cyclotron radius is much smaller than the typical dimensions of the cross. It is not difficult to see that the only electrons that can now contribute to the current (both total end net) are the "skipping orbits" introduced in section 2.2 (see e.g. figure 2.4). These trajectories adhere to the walls of the system while experiencing specular reflections. Figure 3.11 shows the resulting electron paths. The most striking feature of the resulting electron trajectories is that we very simply can evaluate the transmission probabilities in this case. As an exemple, an electron leaving contact reservoir 2 will be "pressed" against the lower wall, bouncing its way to contact 3. This will be so for *each electron* from contact 2, or T_{32}=1. This immediately implies that the electrochemical potential of contact 3 will become *identical* to that of contact 2. Similar arguments hold for contacts 3=>4, 4=>1 and 1=>2, with all associated T's being unity.
It is important to note that in contrast electrons leaving from contact 3 will *not* arrive at contact 2 *directly*. This implies T_{23}=0 and consequently we find $T_{32} \neq T_{23}$. Comparing this to the zero magnetic field case, it is not difficult to see (e.g. from figure 3.9b, zero field traces a1 and b1) that the topological symmetry of the system yields T_{32}=T_{23} at B=0. The large field case of figure 2.11 demonstrates again in a rather striking way the time-reversal symmetry breaking action of the magnetic field. With this quite general statement we close this chapter on classical phenomena in the ballistic regime.

General references to the subject

1. "Quantum Transport in Semiconductor Nanostructures", by C.W.J. Beenakker and H. van Houten, in Solid State Physics, ed. H. Ehrenreich & D. Turnbull, Vol 44 (Academic Press Inc, Boston, 1991), sections 12, 14 and 16
2. "Electronic Transport in Mesoscopic Systems", S. Datta (Cambridge University Press, Cambridge, 1995), chapters 1 and 2
3. Textbooks on statistical physics:
"Elementary Statistical Physics", C. Kittel (John Wiley & Sons, New York, 1967);
"Statistical Physics I", Vol 5 of Landau & Lifshitz (Pergamon Press, Oxford, 1980)

4. Quantum ballistic transport

Keywords: Landauer formalism, electron waveguides, modes, conductance quantization, Landau levels, edge states, (integer and fractional) quantum Hall effect

4.1 Introduction

In chapter 3 we have discussed the characteristic properties of electrons in ballistic or clean systems. Throughout that chapter we assumed that the entities contributing to the transport of charge through the system behave classically. This implied that the electrons could be taken as particles following trajectories described by the laws of classical mechanics.
In the present chapter we again focus on the ballistic regime. In this case however the wave nature of the electrons is included and we will see how this affects the transport in mesoscopic-sized systems.
Following the short introduction of this section (4.1), in section 4.2 we discuss the conductance in a ballistic system in terms of an elastic wave scattering problem and derive the famous Landauer expression for it, which provides a simple relation between the transmission probability and the conductance between two points. From this description based on wave propagation the natural role of 1-dimensional transmission modes or quantum channels will become evident. In section 4.3 the effect of a large magnetic field on the properties of electrons is reviewed and the formation of Landau orbits and its consequences for the density-of-states is demonstrated. Next, in section 4.4 the results of the two preceding sections are employed in a 2-dimensional electron system to show the formation of conducting channels at the boundary of the system, denoted edge states or edge channels. These 1-dimensional quantum channels will be used as the foundation for the description of the famous (integer) quantum Hall effect. The chapter will be closed by briefly mentioning a few of the more exotic phenomena in high fields, namely the fractional quantum Hall variant and the recently introduced Composite Fermion; this is to be found in section 4.5.

4.2 Electrical conductance, transmission and conductance quantisation

In this section we will introduce the key aspects of the description of the transport of electrons in the ballistic regime. Elaborating on a brief introduction given in section 2 of chapter 3 we will show how the conductance of electrons in a clean system is simply determined by the probability of the particles to reach the appropriate contacts. In addition, the ballistic properties of the system combined with the wave (or quantum) nature of the electrons will allow us to derive a quantitative relation between this probability and the conductance, leading to the famous Landauer expression for the conductance. From this derivation the importance of the 1-dimensionality of the contributing electronic quantum states or channels will become clear, in particular as we will show that each quantum channel yields exactly the same contribution to the conductance of the system. The most striking consequence of this "equal contribution" aspect is that the conductance of a finite-width system is quantised in units $2e^2/h$.

In section 3.2 of the preceding chapter we discussed the classical transport of electrons in mesoscopic systems in the ballistic regime; it was limited to the low-bias or linear regime, i.e. assuming all applied potentials in the system to be much smaller than the Fermi energy (or more precisely, E_F/e). From that discussion, centered around the figures 3.1, 3.2 and 3.5, three main notions emerged. First, the role of electric fields (or gradients of the potential) resulting from the small potential differences applied to the contacts in the system was shown to be negligible. This is a direct consequence of the fact that in the ballistic regime and under linear conditions the relevant particles travel close to the Fermi velocity, while variations in the velocity resulting from (local) electric fields will be much smaller. It is this consequence which allowed us to use (classical) ballistic laws. Note that this fact leads to a tremendous simplification of the problem, as we do not need to know how the electric fields are distributed throughout the system. In addition, note the important consequence of this point that the *local* velocity does not depend linearly on the *local* electric field, and so the transport is intrinsically non-local! This should be contrasted to the common, diffusive case where, for isotropic conductors, we always assume the local current (or velocity) and local electric field to point into the same direction, related via the (scalar) conductivity.
Secondly, the ballistic nature of the transport allows the direct evaluation of the probability of an electron that emerges from one contact k to reach a second contact j. From the geometry of the system, including its walls etc., the transmission probability T_{jk} can be calculated. This is particularly simple in the classical case of chapter 3, but also in the present quantum case this can be done in principle. As the probability is a direct measure of the fraction of (a) particle(s) that, emitted from one reservoir and arrives at the second reservoir, it allows us to quantify the distribution of the (net and total) currents over the various contacts of the system. From this distribution of the currents the associated electrochemical potentials (or voltage differences) can be evaluated. Below we will do this quantitatively for a simple strip-shaped structure.
Third, the fact that the excess electric fields within the system resulting from the applied potentials do "not" affect the flow of the particles implies that, within the actual system of interest, there is no dissipation of kinetic energy of the electrons to the rest of the system. This dissipation only occurs within the reservoirs themselves (where full thermal equilibration of all particles is assumed to take place).
With having rehearsed/recapitulated these three important basic points for ballistic systems we are now in the position to employ these to the quantum ballistic case.

As introduced in chapter 1 in the quantum ballistic regime the various length scales are related to each other as

$$\lambda_F, L \; < \; l_e \; < l_i$$

i.e. the Fermi wavelength and the system size are both smaller than the elastic and inelastic mean free path. The fact that λ_F and L may have comparable values allows the occurrence of interference effects in the system, in much the same way as one finds in optical phenomena such as single-slit experiments. Having l_i larger than any of the other relevant length scales assures that the phase of the waves is preserved, a prerequisite for any interference effects to take place. The condition for l_e assures that we are in the ballistic regime.
Figure 4.1 shows a typical ballistic system with a tube- or strip-shaped conductor of width W in the x-direction and length L in the y-direction, connecting two electron reservoirs. Following the

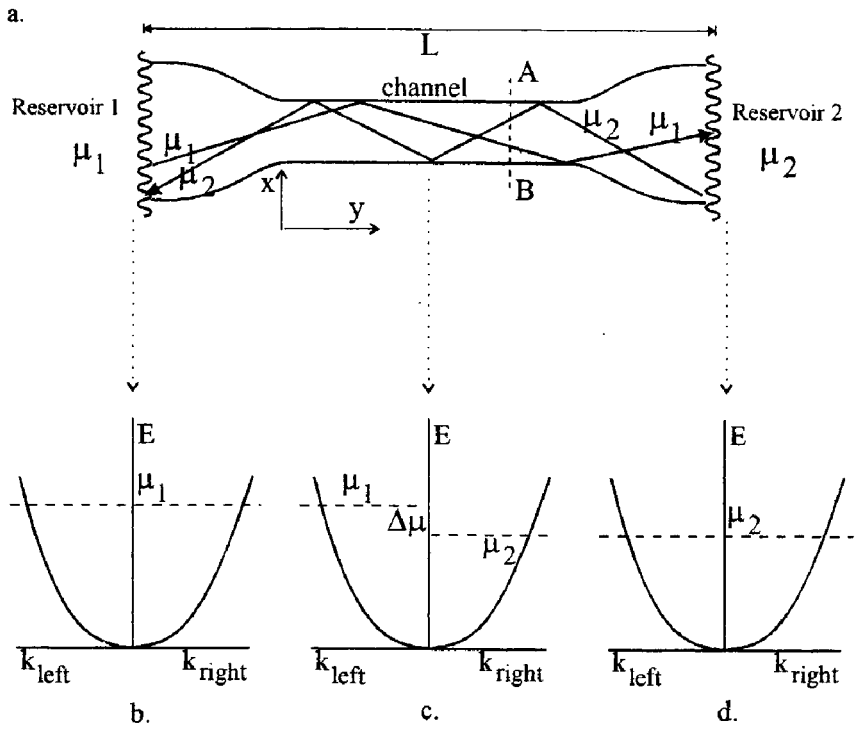


Figure 4.1. Electrons in a ballistic channel. a. The 2D channel of length L (in the y-direction) and width W (in the x-direction) is connected to two reservoirs characterised by electrochemical potentials μ_1 and μ_2. b. The E-k dispersion relations for electrons: b and d in the two reservoirs, and c within the channel. Note the difference in occupation in the channel for the right-going states (filled up to μ_1 from the left reservoir) and the left-going states (filled to μ_2 by the right reservoir).

common use in the literature the interconnecting central strip is denoted as the channel: it should be well-distinguished from the notion of a quantum channel or mode to be defined below! The electron reservoirs or contacts are characteriised by their electrochemical potentials μ_1 and μ_2. This difference in electrochemical potential μ_1-μ_2=$\Delta\mu$=-eV_{12} results from an externally applied voltage between the two contacts 1 and 2, V_{12}. Taking the whole system to be at zero temperature, the electrochemical potential defines the maximum energy up to which the electron states are occupied. For the sake of simplicity it is assumed that the strip or channel is 2-dimensional (2D); we will return to this point at the end of this section.

From figure 4.1a it is obvious that the right reservoir can *only* be reached by electrons with a *k*-vector component pointing towards the right, i.e. right-going *k*-states moving in the positive *y*-direction; similarly the opposite holds for left-going states moving in the negative *y*-direction that can arrive at the left reservoir. As we are interested in the quantum properties of the electron transport we have to describe it in terms of electron states associated with electron waves. Let us focus on the quantum states within the channel region, say at the intersect indicated by the dashed line A-B, to see how these can be described, including their occupation via the reservoirs. In the *y*-

direction the electrons are completely free to travel. This implies that for k_y the electron states are given by the parabolic free electron relation; this is shown in figure 4.1b, c and d. Clearly the states at the left contact are filled up to the (higher) electrochemical potential μ_1 (figure 4.1b) and those at the right contact to the (lower) value μ_2 (figure 4.1d). In the channel region right-going states only are filled from the left reservoir and so will be occupied up to μ_1, as shown in figure 4.1c. Similar arguments hold for the left-going states. As the right-going states are filled up to a higher energy, it is clear that more (and also faster) electrons arrive at the right contact than will be emitted from it, yielding a net particle current from right to left (or a net charge current from left to right).

Before proceeding to evaluate the current we have to take into account also the transverse momenta k_x of the electrons. For this direction the electrons are confined and can not move freely. As discussed in chapter 1 this results in the formation of bound states, the number of states depending on the width of the channel. Figure 4.2 shows this schematically, assuming the confinement to have vertical walls (in E-x-space). The resulting eigenstates have associated energies E_{nx}, which for an (infinitely high wall) box-shaped confinement can be calculated readily (see chapter 1). With the k_x quantised and denoted by the index n the total energy of an electron residing in a state characterised by $\{n,k_{ny}\}$ is given by

$$E_n(k) = E_{nx} + \frac{\hbar^2}{2m} k_{ny}^2 \qquad (4.1)$$

Electrons residing in a state with transverse quantum number n are said to occupy the n-th *mode* or *subband* or *quantum channel* that is available for transport in the channel (note here the different meaning of the word "channel"!). Note that at a given total (kinetic) energy for a high mode index n the kinetic energy in the y-direction is small, as a large n implies E_{nx} to be (relatively) large.

Now we make an important step. As each electron is characterised by the quantum numbers n and k_{ny}, with the first being discrete and the second continuous, they simply behave *1-dimensional*. It is this 1D nature of the electron states which allows us to very elegantly evaluate the net current through the channel. We first evaluate the total and net particle currents, which via the charge-per-particle directly yields the (charge) current.

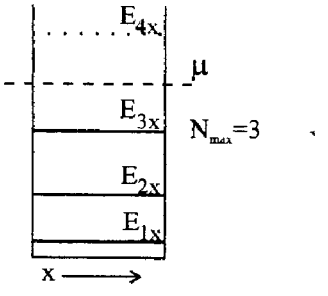


Figure 4.2. The bound states resulting from the confinement of electrons in the transverse (or x-) direction of the channel of figure 4.1a. The successive eigenstates with eigenenergies E_{nx} are indexed by the mode number n. The highest mode that is occupied (i.e. below the electrochemical potential μ) is denoted by N_{max}

The particle current $J_{p,n}$ that flows within a single mode n is simply given by the product of the density and the velocity of the particles, integrated over energy,

$$J_{p,n} = \int_0^{\mu} v_n(E) * \rho_{1D,n}(E) dE \tag{4.2a}$$

with v_n denoting the particle velocity and $\rho_{1D,n}$ being the *1-dimensional density-of-states* for mode n; both these quantities are energy dependent. The *total* particle current or number of particles flowing through the channel in one direction thus follows by summing the contributions of all modes n

$$J_p(\mu) = \sum_n^{N_{max}} J_{p,n}(\mu) = \sum_n^{N_{max}} \int_0^{\mu} v_n(E) * \rho_{1D,n}(E) dE \tag{4.2b}$$

N_{max} denotes the highest occupied mode, i.e. for which the bottom (with $v_{Nmax}=0$ and $k_{N_{max}} (\equiv k_{N_{max}y}) = 0$) is still below the electrochemical potential of the system (figure 4.2). From figure 4.1 it is obvious that the *net* particle current through the channel (say at A-B in figure 4.1a) simply is the difference between the right-moving particles (in states up to μ_1) and left-moving particles (up to μ_2, with $\mu_1 > \mu_2$), i.e.

$$J_{p,net} = J_{p,right}(\mu_1) - J_{p,left}(\mu_2) \tag{4.3}$$

From quantum mechanics we know that the velocity of a particle is given by

$$v_n(E) = \frac{1}{\hbar} \frac{dE_n(k)}{dk} \tag{4.4a}$$

For the *1-dimensional case* the density-of-states (disregarding the spin) is given by (see chapter 1)

$$\rho_{1D,n}(E) = \frac{1}{2\pi} \left(\frac{dE_n(k)}{dk}\right)^{-1} \tag{4.4b}$$

which makes the product of these two factors

$$v_n(E) * \rho_{1D,n}(E) = \frac{1}{h} \tag{4.5}$$

i.e. it is *independent of energy E or mode index n*, and given by (the inverse of) a constant of nature, *h*. Taking into account that each state can accommodate $g_s=2$ electrons (due to the two spin directions) e.q (4.5) should be multiplied by this factor.
By combining eqs. (4.2)-(4.4) the net particle current contribution for *one single* mode becomes

$$J_{p,n,net} = J_{p,n,right}(\mu_1) - J_{p,n,left}(\mu_2) = g_s \frac{\mu_1 - \mu_2}{h} = \frac{2\Delta\mu}{h} \tag{4.6}$$

Remember that $\Delta\mu = -eV_{12}$ is the difference in electrochemical potential resulting from the applied bias voltage between the two contacts 1 and 2. Noting that each particle has charge $q=-e$ we thus find for the *net charge* which flows from left to right in (the single) mode n

$$J_{e,n,net} = g_s * -e * J_{p,n,net} = -2e \frac{\Delta\mu}{h} \tag{4.7}$$

and so for the *total net* current as the sum over all contributions per mode

$$J \equiv J_{e,net} = \sum_{n=1}^{N_{max}} J_{e,n,net} = N_{max} * -2e \frac{\Delta\mu}{h} = N_{max} * \frac{2e^2}{h} * V_{12} \tag{4.8}$$

This can be rewritten in terms of a (two-terminal) conductance, reading

$$G_{2t} \equiv G_{12} \equiv \frac{J}{V_{12}} = N_{\max} * \frac{2e^2}{h} \tag{4.9}$$

This is the famous expression for the two-terminal conductance of a mesoscopic ballistic system. It expresses that this conductance is *universal*, as it is given in units of the *conductance quantum* $2e^2/h$ which only contains constants of nature, and that it is *quantised* as it scales with the *integer number* of occupied quantum channels. Note that the factor of two directly follows from the spin degeneracy. The universality is a direct consequence of the cancellation of the energy dependence of the velocity and the density-of-states, which *only* holds for 1-dimensional systems (see eq. (4.5)). This leads to the important result that *each* mode n contributes *exactly* the same amount of current to the total net current, as can be seen directly from eq. (4.8).
From figure 4.2 and from chapter 1 we see that the transverse confinement due to the width W determines the number $N_{\max}$ of subbands or modes with an eigenenergy E_{nx} below the Fermi energy (or electrochemical potential) and that will occupied. Assuming again that the confinement is square-well shaped, we can estimate this number by noting that

$$k_{x,n} = n\frac{\pi}{W} \tag{4.10a}$$

and

$$k_{x,n}^2 + k_{y,n}^2 = k_F^2 \tag{4.10b}$$

The value $n=N_{\max}$ if obtained from the condition $k_{x,N_{\max}} \le k_F \le k_{x,N_{\max}+1}$. Assuming that a reasonable number of quantum channels is transmitted we have $N_{\max} >> 1$ and so it is clear that for most n we will have $k_{y,N\max} << k_F$ and so $k_{x,N_{\max}} \approx k_F$, which implies (from eq. (4.10a))

$$W \approx N_{\max}\frac{\pi}{k_F} = N_{\max}\frac{\pi}{2\pi/\lambda_F} = N_{\max}\frac{\lambda_F}{2} \tag{4.11}$$

So, changing the width of a structure allows one to tune the number of 1D modes that is transmitted through it. As eq. (4.9) shows that the conductance depends linearly on the number of transmitted modes, $N_{\max}$, we conclude that the conductance of a (narrow) opening will increase *stepwise* with increasing width: each time the width increased so much as to allow the next mode to become occupied and transmitted the conductance will show a stepwise increase by the universal amount $2e^2/h$. Note that the stepsize will be only e^2/h if each state will be occupied by a single spin direction only, e.g. in the case of a large magnetic field. We will return to this aspect in the next section of this chapter.

Before proceeding to generalise the Landauer formula we first want to show some experimental evidence for the rather remarkable conductance quantisation. To this purpose a narrow constriction or *point contact* is required, of the same nature as the one that was introduced in chapter 3 around figure 3.6 and 3.7a. Starting from a nearly ideal 2-dimensional electron gas at the interface between two III-IV materials with a large elastic mean free path (from chapter 1 we know l_e can be tens of μm's) properly shaped gates are realised at the surface by e-beam lithography. The configuration is shown schematically in figure 4.3. With a negative voltage to the gates the electrons underneath it will be repelled, leaving a short and narrow channel for the electrons to travel from the one half-plane of 2DEG to the other. In this way the shape of the gate pattern is "imprinted" into the 2DEG, defining the narrow channel or point contact required for

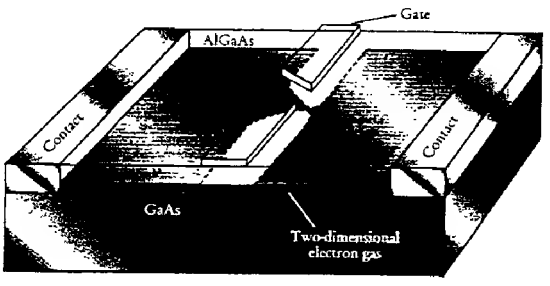


Figure 4.3. A cross section of a (quantum) point contact or QPC realised in a 2DEG. Surface gates with a negative voltage relative to the 2DEG remove the electrons of the 2DEG underneath, imprinting a short and narrow channel in the 2DEG. The width W of the channel can be controlled by the gate voltage.

our experiment. Note that the width W of the channel can be controlled by the gate voltage, with a more negative value making it narrower. The conductance through the point contact is measured via two contacts attached to the two large 2DEG areas.
The results at low temperature (~ 1K) are shown in figure 4.4, as obtained by B.J. van Wees et.al. (Phys. Rev. Lett. 60, 848 (1988)). The behaviour of the conductance fully complies to the "predictions" made above, i.e. a nice stepwise increase with increasing (i.e. less negative) gate voltage or increasing width W of the point contact, with step size $2e^2/h$. In the original experiment up to 16 steps were seen. Because of the quantised behaviour of the conductance of such a contact it is commonly referred to as a *quantum point contact* or *QPC*. It may be of some (historical) interest to mention that in this case the experimental result *preceded* the theoretical description.

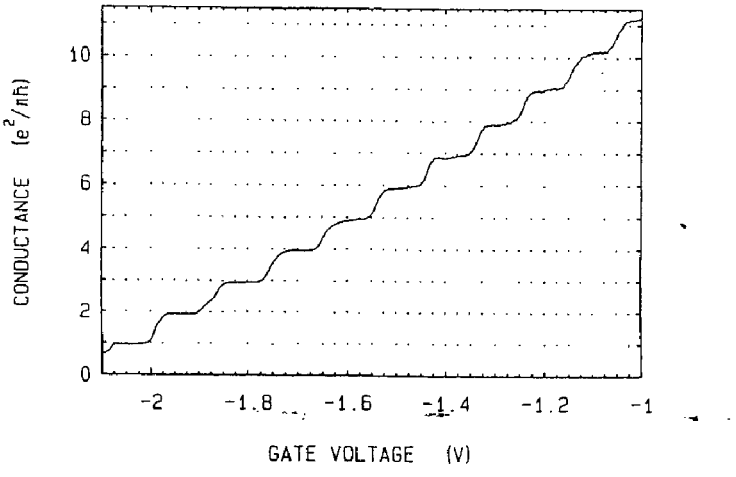


Figure 4.4. The quantisation of the conductance in a (quantum) point contact of variable width, realised by a gated structure on a GaAs-AlGaAs heterostructure containing a high mobility 2DEG; less negative gatevoltage results in an increase in width. It shows the stepwise increase in units $2e^2/h$ for each additional mode that is transmitted through the QPC.

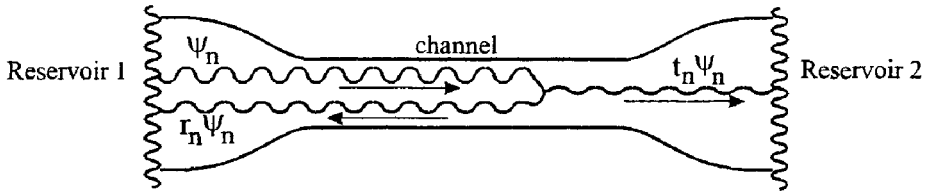


Figure 4.5. Schematic representation of a state in a certain mode n with a wavefunction $\psi_{n,ky}$ that is partially reflected inside the system, leading to an effective transmission probability T_n.

It is important to note that in the derivation of eq. (4.9) it has been assumed that each particle/wave emitted from (say) the left contact in a mode that is allowed in the center of the channel, will reach the right reservoir. Stated differently, we require a complete or *unity transmission* through the channel for each wave in an allowed mode. The natural question arises of what will occur if this condition is relaxed, and we allow waves to be transmitted only partially. Figure 4.5 schematically shows how the wave can be represented in the system.
Let the wavefunction $\psi_{n,ky}$ represent the wave in mode n and let us assume that its transfer from the left to the right reservoir is characterised by a (complex) transmission coefficient t_n, i.e.

the transmitted partial wave is given by $t_n \cdot \psi_{n,ky}$ (4.11a)

and the reflected partial wave by $r_n \cdot \psi_{n,ky}$ (4.11b)

Taking the wave function to be normalised, this implies for the transmission coefficient t_n and the reflection coefficient r_n

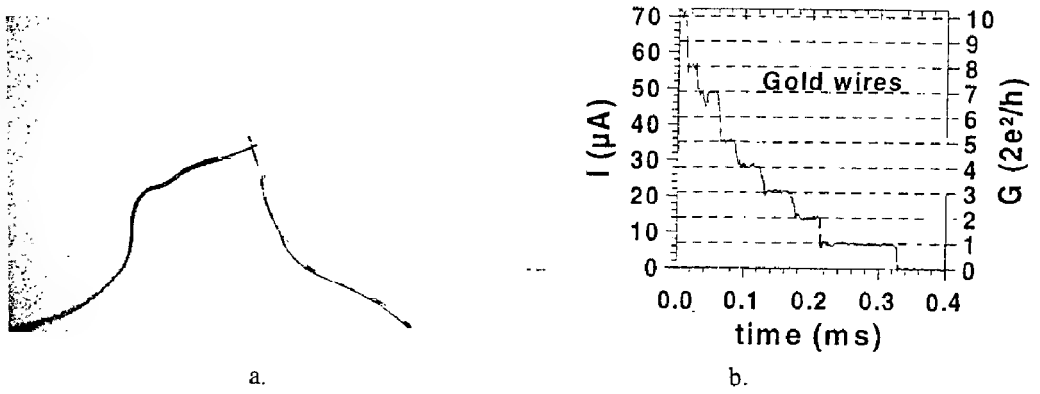
$$t_n^* t_n + r_n^* r_n = |t_n|^2 + |r_n|^2 = T_n + R_n = 1 \qquad (4.11c)$$

Here T_n is called the *transmission probability* and similarly R_n is denoted as the *reflection probability*. Note that the normalisation (4.11c) is the same as that we require the *conservation of particles* or of *wave probability* to hold! As now the probability for the wave to reach the other side of the system (i.e., to be transmitted) is given by $T_n \leq 1$ it is not difficult to see that the contribution to the current for the n-th mode in eqs. (4.2) and (4.8) is reduced by the same factor. This implies that the conductance can be written as

$$G_{12} \equiv \frac{\sum J_{e,n}}{V_{12}} = \frac{2e^2}{h} \sum_{n=1}^{N_{max}} T_n \qquad (4.12)$$

This expression is one representation of the famous *Landauer formula* for the conductance of a mesoscopic system. Note that if we take T_n=1 for n up to N_{max} and T_n=0 otherwise we regain the original expression eq. (4.9). So, a finite transmission probability leads to a reduction of the step size, which will break down the property of universality of the conductance quantisation.

Before closing this section we want to return to the simplification that was made just following figure 4.1, taking a 2-dimensional strip as our starting point for the channel. Evidently, the whole discussion was centered around the further reduction of the dimensionality of the system to 1D, as only for that specific case the energy-dependence cancellation of the velocity and the density-of-states (eq. (4.5)) holds. By the same token, if we would have started from a 3D structure, the wire



a. b.

Figure 4.6. Quantization of the conductance of a metal-metal contact, formed by two touching metal wires. By vibrating the two wires relative to each other the contact conductance is varied in time, which is measured. a. shows the wires, with their bare ends bent such as to weakly touch one another. b. displays the result obtained for two gold wires. It clearly shows the plateaus in the conductance at integer units $2e^2/h$.

should be made so narrow in two directions that all transport takes place via a limited number of 1D quantum channels. If this is so the wire will show quantised conductance.
A recent very simple and in this way highly instructive experiment shows the validity of this argument, performed by J.L. Costa-Kraemer et.al. (Surface Science 342, L1144-L1149 (1995)). It basically consists of two macroscopic metallic wires (of a few tenths of mm diameter) that are bent such as to touch each other loosely (see figure 4.6a), in this way forming a small-sized metal-metal contact. The main idea now is that the conductance of the contact will depend on the force pressing the two wires to one another. So, by varying this contact force the conductance will vary as well, and it would be of interest to see whether the conductance indeed shows quantization under these conditions. The conductance is measured by applying a fixed bias voltage of a few tens of mV and measure the current flowing through the contact. The force on the contact is varied by just "tapping" onto the fixture that holds the two wires. This leads to vibrations of the wires, yielding the required variations of the force in time. A trace of conductance-versus-time thus should show the anticipated phenomenon. This is indeed the case, as is seen in figure 4.6b. Here two gold wires are employed. The conductance clearly shows quantization in units $2e^2/h$, just as predicted.

Before closing this section we want to present one more experiment relating to the phenomenon of quantised conductance. Evidently, the quantised nature of the conductance results from the stepwise change of the number of modes that can be transmitted through the channel if its width is varied. Basically no restriction on the type of waves is posed in the derivation given above (except the cancellation of eq. (4.5)), and so a similar phenomenon should be obtainable for acoustical or optical waves. This has indeed been shown in an experiment by E.A. Montie et.al. (Nature 350, April 18 (1991)) using near infrared optical waves of λ=1.55 μm. Figure 4.7a shows the set up of the experiment. Here the light is allowed to enter a detector volume ("integrating sphere" with

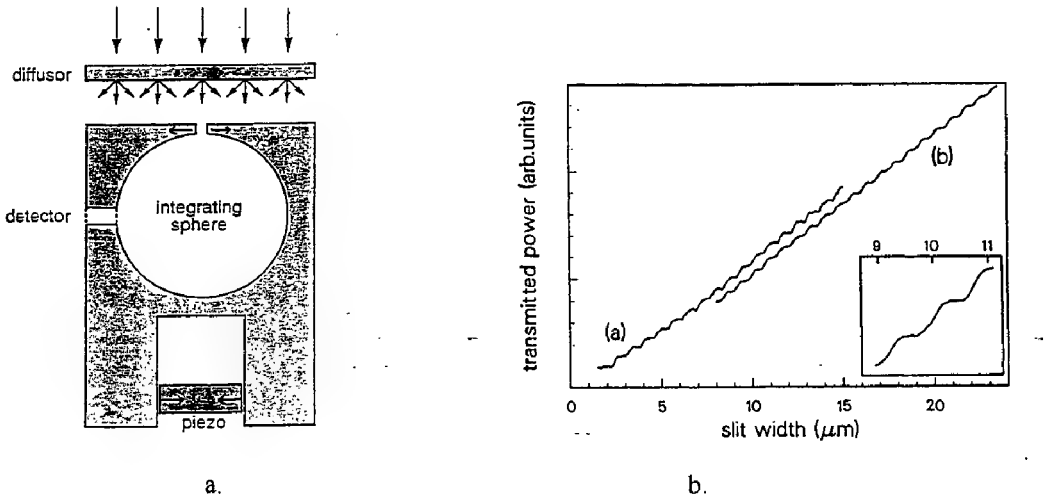


Figure 4.7 The optical analogue of the conductance quantisation. a. schematically shows the apparatus, with the diffuser used to uniformly occupy all optical modes, and the integrating sphere employed to measure the true total transmitted power. b. shows the result of the transmitted power (or the total transmittance) in dependence of the width of the slit: here a and b represent two different measurement runs.

with added detector) via a slit the width of which can be varied via a piezo driving element. Note that the slit decreases in width if the piezo element is forced to expand (via a voltage applied to the ends of the element). Figure 4.7b shows the result. We see a clear stepwise increase of the transmitted power or the total transmittance through the slit in dependence of its width. Note that each step requires the slit width to increase by ~ 0.7 μm or half a wave length, which is in accordance with the fact that one additional mode is allowed (cmp. to eq. (4.11)).

Problem: One important difference between the optical and the electron interference experiment is the absence (for the optical case) of the energy-dependence cancellation, as seen to exist for the eleoctron case (eq. (4.5)). Discuss the consequences of this difference.

In the section we have limited the discussion to quantum ballistic transport in zero magnetic field. In the next two sections the effect of magnetic fields will be discussed, and in particular the large-field regime will be investigated.

4.3 Landau quantisation in large magnetic field

At a number of places before we have discussed the effect of a magnetic field on the motion of electrons. In chapter 1 it was shown that the linear motion of a charged particle like an electron becomes curved due to the Lorentz force acting on the charge. The resulting orbit is circular perpendicular to the applied field B (see fig 1.9) with a cyclotron radius depending on the velocity

of the particle and the field strength (eq. (1.8c), encircled with the angular cyclotron frequency given by $\omega_c = eB/m$ (eq. (1.8b)). Note that the cyclotron frequency is *independent of the velocity* of the particle, so all electrons revolve at the same rate! This motion, which is completely *classical*, results in effects like the reduction of the resistance of narrow channels with diffusive wall scattering (see section 2.2, figures 2.4-2.7), the occurrence of electron focusing (section 3.3, figures 3.6 and 3.7) and the formation of skipping orbits as introduced in section 2.2 (figure 2.4) and used in the Hall effect discussion at figure 3.11.
However, in addition to the classical effect of introducing a curvature of the path of the charged particle we also introduced a *quantum* effect on the field on the electron trajectory by realising that the canonical momentum p is modified via the vector potential A associated with the field strength B; this was done in section 2.3 in relation to the effect of the magnetic field on the *quantum* phenomenon of Weak Localisation.
In this section we will study the quantum effects resulting from the interaction of electrons with a large magnetic field. In particular we will see how the field leads to the formation of quantisation of the areal size of the electron orbits and how this modifies the density-of-states from its zero-field (quasi-) continuos character to a fully discrete spectrum, each state being associated with a so called Landau level.

As already introduced in section 2.2 it can be shown that a magnetic field B modifies the canonical momentum p via the vector potential A defining the field via $\vec{B} = \vec{\nabla} \wedge \vec{A}$ following

$$\vec{p} = \hbar\vec{k} = m\vec{v} - q\vec{A} = m\vec{v} + e\vec{A} \qquad (4.13)$$

Note that the quantum mechanical aspect only comes in via the relation $p = \hbar k$, as the remainder of eq. (4.13) directly follows from classical mechanics and electrodynamics. The main quantum consequence is that the electron wave vector (and so the wave length and the phase) now is affected by the vector potential or the field. As the magnetic field only couples to the directions that are perpendicular to the field, we limit our discussion to the case that our system is 2-dimensional and that the field is taken perpendicular to the 2D plane. In addition we assume an ideal, ballistic system without scattering.
Already from the classical cyclotron motion which implies a harmonic oscillator behaviour, the application of quantum mechanics should lead to the formation of eigenstates with energies quantised in units associated with the classical oscillating frequency. For the Hamiltonian we can write

$$H = \frac{(P + eA)^2}{2m} + U(x,y) \qquad (4.14)$$

with P denoting the momentum operator and $U(x,y)$ representing the electrostatic potential in which the electrons move. We first assume a uniform, unbounded 2D system, and so we may take $U(x,y)=0$ everywhere. Introducing a particular gauge for the vector potential, $\vec{A} = (0, Bx, 0)$ then elementary quantum mechanics yields the following eigenvalues for the energy of the electrons

$$E_n = (n + \frac{1}{2})\hbar\omega_c \qquad (4.15)$$

with ω_c being the (classical) cyclotron (angular) frequency and n=0,1,2,3,... being the index that classifies the particular "state". These "states" are called Landau levels. So, at (large) magnetic fields electrons are only allowed to occupy these levels, or, stated differently, the zero-field *continuous* 2D density-of-states (see eq. (1.16b)) is converted into a *discrete* DOS at high field.

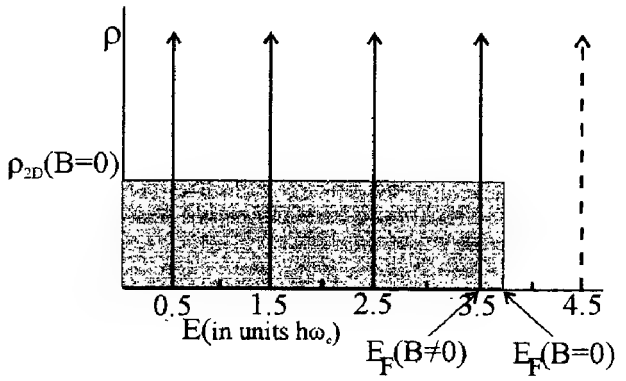


Figure 4.8. The densities-of-states $\rho(E)$ for zero and high magnetic field. At zero field all states in the constant 2D zero field DOS are filled up to the (zero field) Fermi energy $E_F(B=0)$, with states beyond this energy being empty. At large fields the DOS becomes discrete forming Landau levels (arrows) at energies $(n+1/2)\hbar\omega_c$.

Note that this implies that now *many* electrons occupy a "state" with the *same* energy, i.e. such a "state" is strongly *degenerate*, reason for us to denote these entities by levels. At first sight this may seem to lead to a rather fundamental problem related to the fermion nature of the electrons involved. From the Pauli exclusion principle we know that no two fermions are allowed to occupy the same quantum state. However, realising oneself that not all electrons occupy the same $\{r=x,y\}$ position in the 2D system, the added quantum number associated with position again leads to individual quantum numbers $\{n,r\}$ for each electron.
Now let us determine the degree of degeneracy per Landau level. Although this can be done exactly we want to employ a simple picture for this purpose, based on the conservation of the total number of electrons in the system. Figure 4.8 shows the DOS $\rho(E)$. For the zero-field case the DOS for 2D is constant. At zero temperature the states are occupied up to the zero-field Fermi energy $E_F(B=0)$, indicated by the filled rectangle, and empty for all states at energies beyond. The high field DOS is denoted by the δ-functions (arrows) positioned at the energies given by eq. (4.15). Assuming that the total number of states is the same for zero field as well as for high field (which is reasonable, as the total number of electrons will not change as well!) we immediately see that all electrons residing in an energy interval $\hbar\omega_c$ will "condense" into a single Landau level as follows: states from 0-$\hbar\omega_c$ enter the n=0 state at $E_0 = (1/2)\hbar\omega_c$, states between $\hbar\omega_c$- $2\hbar\omega_c$ condense into E_1, etc. From eq. (1.16b) for the 2D case the number of states *for an area S and per unit energy* is given by

$$N_{2D}(B=0) = \rho_{2D}(B=0)*S = g_s\frac{m^*}{2\pi\hbar^2}*S \qquad (4.16)$$

with g_s=2 for the spin-degeneracy. Thus the number of states available in a single Landau state and within the area S becomes

$$N_{L,n} = N_{2D}(B=0)*\hbar\omega_c = g_s\frac{m^*}{2\pi\hbar^2}*S*\hbar\omega_c = g_s\frac{e}{h}BS = g_s\frac{e}{h}\Phi = g_s\frac{\Phi}{\Phi_0} \qquad (4.17)$$

Here Φ denotes the total flux $B*S$ penetrating the area S and $\Phi_0=h/e$ is the single-electron *flux quantum*. By taking a unit area equation (4.17) immediately provides the number of states per Landau level or the "DOS" for this case. One word of caution concerning the use of the "DOS" of eq. (4.17) might be useful, as it is not a quantity given per unit energy but per "unit level" in stead! Note that eq. (4.17) states that the number of electrons which can be accommodated (in a given area) in a Landau level $\{n\}$ increases linearly with the applied magnetic field.
From eq. (4.17) we can derive a convenient quantity relating the number of electrons in the area S, $N_e=n_eS$, to the number of flux quanta $N_\Phi=\Phi/\Phi_0$ penetrating the same area, called the filling factor

$$\nu = \frac{N_e}{N_\Phi} = \frac{n_e S}{\Phi/\Phi_0} = \frac{n_e \Phi_0}{B} = \frac{n_e h}{eB} \tag{4.18}$$

It is important to note that the formation of the δ-functions for the non-zero field DOS as shown in figure 4.8 imply that *many* electrons occupy states with exactly the *same* energy. This may seem rather striking in view of the fact that electrons are Fermions and so they should comply with Pauli's exclusion principle which states that no two Fermions are allowed to occupy a single quantum state. This implies that all the electrons occupying the energy-degenerate states in a single Landau level should have at least one quantum number that is unique to each individual particle. It is not difficult to see that this is the position vector $\{r\}=\{x,y\}$, i.c. individual orbiting electrons residing in the same Landau level try to prevent mutual overlap. Although maybe superfluous to say, this is not required for electrons in different Landau levels: here the different Landau level index n makes up the distinction in quantum numbers and so these orbiting particles are allowed to occupy the "same position" in space!

Expression (4.17) allows us to calculate the area δS in the $\{x,y\}$ plane occupied by a single (spin-degenerate) state, by simply taking $N_{L,n}=g_s=2$, which yields

$$\delta S = \frac{h}{eB} = \frac{\Phi_0}{B} \tag{4.19}$$

This leads to the important conclusion that *each spin-degenerate state in each Landau level occupies such an area in real space that it contains exactly one single fluxquantum.* Realising that this area is "shared" by two electrons (spin-up and spin-down) the area per electron is only half that given by eq. (4.19), and so one can associate a length scale (or radius) with a single electron, given by

$$l_B = \sqrt{\frac{\delta S}{g_s \pi}} = \sqrt{\frac{\hbar}{eB}} \tag{4.20}$$

This new quantum *length scale* is called the *magnetic length.*
It is important to note that the preceding conclusion does *not* imply that the (real space) spatial extent of Landau level states of *different* quantum number n is the same. To explain this more clearly we want to see what relation exists between this new quantum length scale and the cyclotron radius as its classical counterpart. From eq. (4.19) we see that the area *covered* by each Landau level state is exactly the same, irrespective of the level quantum number n. Referring to figure 4.8 it is obvious that the lower Landau levels will be occupied by electrons originating from lower-lying states, i.e. by electrons of a smaller kinetic energy. The corresponding "orbits" will have a smaller "radius" as well, and so these will "trace out" a smaller area in space as compared to those associated with the higher lying states. The same thing follows (of course!),

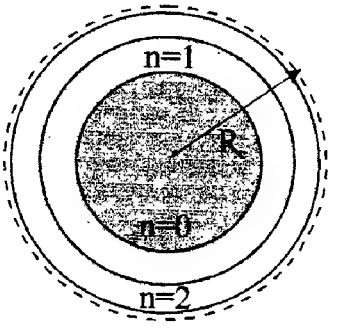


Figure 4.9. The real space areas covered by electron states in a 2-dimensional electron system in a large magnetic field. The resulting Landau levels are identified by the quantum numbers n=0, 1, 2, ... The area for each (disc- or ringshaped) state is the same implying that the area of each state encloses the same flux, namely one fluxquantum Φ_0. The dotted outer circle represents the classical cyclotron radius.

however now in an exact way, by evaluating the solutions to the Hamiltonian (4.14) such as to determine not only the energy eigenvalues but also the eigenfunctions. These also show that the Landau states of high-index number indeed have a larger extent in real space than those from the lower levels. From these Correspondence Principle arguments the following picture arises (see figure 4.9 and 4.10). Figure 4.9 illustrates that each state covers the same area in space, i.e. the n=0 state occupies the central disc of area $\delta S=\Phi_0/B$, the next state (n=1) occupies the ringshaped strip adjacent to the disc, which also has the same area, etc. This can be continued up to the highest occupied level, determined in the way as shown in figure 4.7.

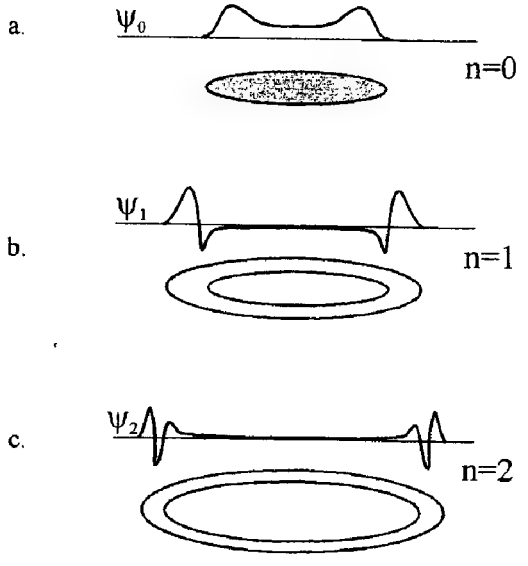


Figure 4.10. Illustration for the approximate shape of the wave-functions for the different Landau level states.
a. shows the area for the n=0 state with a schematic representation of the wavefunction. Note the non-zero value of ψ_0^2 over the area of the disc.
b. shows the similar picture of the next higher state n=1. Note the near-zero probability ψ_1^2 over the center part of the area, demonstrating that the different Landau states try to avoid each other. c. shows the same for the highest lying state n=2.

Denoting N_{max} as the highest occupied Landau level, it is simple to derive the relation between the magnetic length and the classical cyclotron radius $R_{c,Nmax}=mv_{Nmax}/eB$. By equating the kinetic energies one can write directly

$$\frac{1}{2}mv_{N_{max}}^2 = (N_{max}+\frac{1}{2})\hbar\omega_c = (N_{max}+\frac{1}{2})\frac{\hbar eB}{m} \qquad (4.21a)$$

and so

$$R_{c,N_{max}} = \sqrt{\frac{(2N_{max}+1)\hbar}{eB}} = \sqrt{(2N_{max}+1)}l_B \qquad (4.21b)$$

Figure 4.10 illustrates the approximate correspondence between the areas in real space and the shape of the Landau wave functions for the various Landau states. From this combination the correlation between the different areas in space with the electron probability for the different Landau level states becomes clear. Note that the for increasing Landau level quantum number the average orbit size increases. Note also that the distribution of the probabilities over space for the different Landau quantum numbers demonstrates that states associated with different n try to avoid each other, in this way making the different eigenstates of the Hamiltonian orthogonal.

Before closing this section two more aspects may be useful to be mentioned. First, in the discussion of the effect of the magnetic field on the electron orbits and their energies we have completely neglected the effect of the spin of the electrons, except by introducing the spin degeneracy factor g_s=2, e.g. in the DOS (eq. (4.16)). From introductory quantum mechanics we know that the spin will lead to a finite energy difference between a state with spin parallel and anti-parallel with respect to the direction of the applied field. This is called the Zeeman splitting, the size of which is governed by the product of the Bohr magneton μ_B and the Lande g-factor. As a consequence the energy spectrum given by eq. (4.15) will be modified to read

$$E_n(B) = (n+\frac{1}{2})\hbar\omega_c \pm \frac{1}{2}g\mu_B B = (n+\frac{1}{2})\hbar\frac{eB}{m^*} \pm \frac{1}{2}g\frac{eB}{m_0} \qquad (4.22)$$

It is important to notice the difference between the effective electron mass m^* and the free electron mass m_0. For the free electron case the Lande factor is ~ 2, however in semiconductors it may deviate strongly from this value and it even can depend on the magnetic field. Despite this in

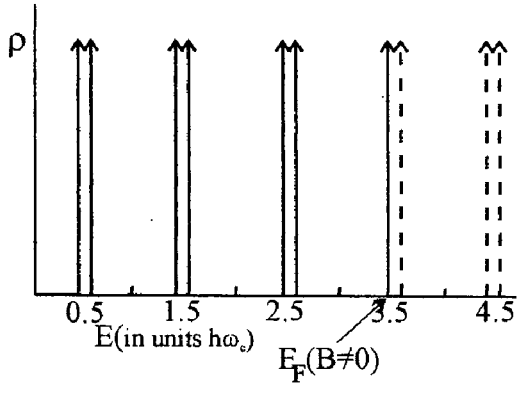


Figure 4.11. The energy spectrum of electrons in a 2DEG in a high magnetic field, including spin. Note that each spin-degenerate Landau level δ-function is resolved due to the Zeeman splitting, yielding separate δ-functions for spin up and spin down states.

semiconductors the spin splitting contribution commonly is more than one order of magnitude smaller than the Landau splitting, mainly because of the strongly reduced effective mass. Figure 4.11 shows schematically the resulting DOS if the spin splitting is included. It should be compared to the arrows shown in figure 4.8.

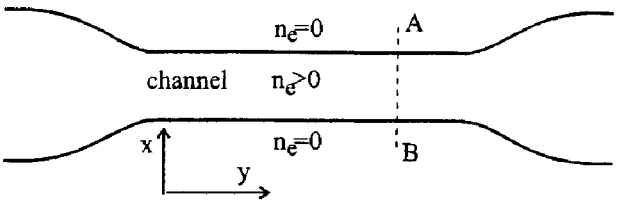
A second aspect that needs mentioning is the notion of the Fermi energy (or the electrochemical potential) for an electron system in a magnetic field. From thermodynamics we know that the Fermi energy is defined as the lowest energy required to add the $(N+1)$st electron to a system containing N electrons. Stated differently, it is the energy of the state immediately beyond the highest occupied state. This clarifies what has been schetched in figures 4.8 and 4.11, where $E_F(B=0)$ and $E_F(B\neq 0)$ are indicated by arrows. While for zero magnetic field the Fermi energy is defined completely by the electron density (eqs. (1.14 and (1.16a)) this no longer is true at non-zero field. The discrete density-of-states, resulting from the formation of Landau levels which contain many states that are degenerate in energy, makes that Fermi energy always lays in one of the Landau levels. As the energy of these levels shifts with the applied field following eq. (4.15 (or (4.21)), the Fermi energy no longer is constant but depends on the field. As many quantities such as the magnetic polarisation, the thermopower, the heat capacity etc. depend on the Fermi energy (and the DOS at the Fermi energy) these will be affected by the magnetic field as well. In the next section we will see that this also holds for the conductance.

Problem: schetch the Fermi energy in dependence of the magnetic field. In particular, consider how it will evolve for such large fields that only the lowest Landau level is occupied. What will be the spin state of the electrons in this last case?

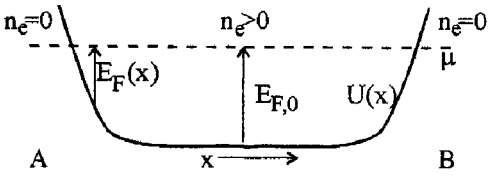
4.4 Edge states and the quantum Hall effect

The preceding section provides all the basic ingredients required to discuss the transport of electrons in large magnetic fields, i.c. in the case Landau levels are formed. Here we will make some connection to the end of Chapter 3, where skipping orbits along the wall of a system were seen to develop if a sufficiently large field was applied. While these skipping orbits were just "simple" classical bouncing orbits, in this section we will see how these develop if quantum mechanics is taken into account. Again, as in the preceding section, we limit ourselves to the 2-dimensional electron gas case.

The derivation which lead to the introduction of Landau levels as the discrete density-of-states for electrons in a large magnetic field started by introducing the Hamiltonian eq. (4.14) for a (single, non-interacting) electron in a magnetic field. To highlight their main aspects it was assumed in section 4.3 that the electrostatic potential energy is constant everywhere in the system, arbitrarily taken as $U(r)$=0. This is equivalent to dealing with a uniform and thus unbounded system. The question to address now is how these quantised Landau level states are affected by the introduction of boundaries or edges in the system. Figure 4.12a shows the typical case of a channel connecting two large 2DEG areas or reservoirs. Note that we have taken the walls or boundaries of the channel to have a finite slope. By definition the channel defines the area



a.



b.

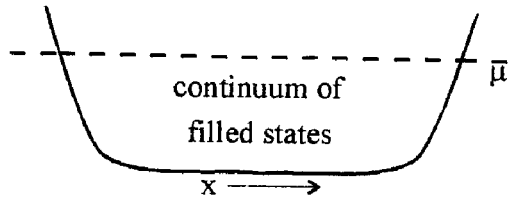
Figure 4.12. A finite width channel in a 2D electron system. In a. the channel is shown in real space; electrons are confined in the x-direction and free to move in the y-direction. Outside the channel the electron density is zero. In b. a cross section of the energy landscape in the (bounded) x-direction is shown. The dashed line is the electrochemical potential μ, which is constant throughout the system. Note the x-dependence of the electron density and so the Fermi energy.

within which conductance takes place and so the electron density is non-zero. This immediately implies that the Fermi energy will be non-zero too. As outside the channel the electron density is zero we arrive at a (x-) dependence of the electron density and the Fermi energy. This is shown in figure 4.12b. It is important to note that, in order to obtain a position-dependent Fermi energy we are not allowed to vary the electrochemical potential. In doing so we would violate the thermodynamic requirement of a constant μ throughout a system of particles that are free to move, which is the case for the electrons inside the channel. As we know that in this area the relation $\mu=E_F(x)+U(x)$ should hold, - i.e. the electrochemical potential is the sum of the kinetic energy and the potential energy - , the channel is actually formed by the position-dependence of the electrostatic potential energy, $U(x)$.

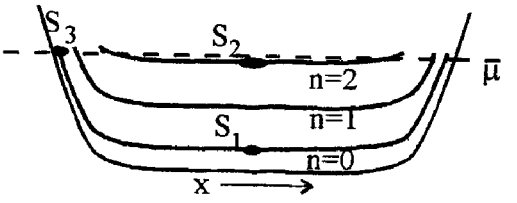
At zero magnetic field the electrons will fill all the states available up to the electrochemical potential (figure 4.13a) (Note: to complicate matters not further we assumed that the channel is so wide that so many transverse modes (see discussion in section 4.2) are formed that actually a continuum of states results). Now the magnetic field is switched on and Landau levels are formed at (kinetic) energies given by eq. (4.15) (neglecting the spin for the moment). For the total energy of an individual state in a particular Landau level it is not difficult to see that we can write

$$E_n(x,k_y)=(n+\frac{1}{2})\hbar\omega_c+U(x)+\frac{\hbar^2k_y^2}{2m} \qquad (4.23)$$

For the time being we assume the last term, representing the kinetic energy in the y-direction which is somtimes called the guiding energy, to be negligible, an assumption we will return to below. Also below it will become clear why we do not include a kinetic term for the x-direction.



a.

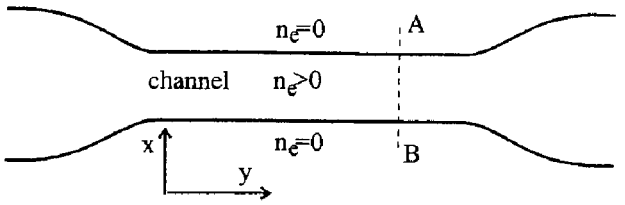


b.

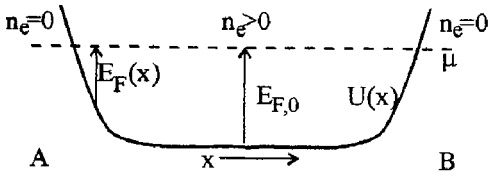
Figure 4.13. Electron states in a channel. a. represents the case for zero magnetic field, where states are formed characterised by the wave vectors $\{k_x,k_y\}$. In b. the magnetic field is switched on and Landau levels are formed, characterised by the level indices n=0, 1, 2, ... Note the three different electron states S_1 at the lowest (n=0) Landau level in the center of the channel, state S_2 at the same position but at the highest occupied Landau level (n=2), and state S_3 in the lowest Landau level but at one of the edges.

Eq. (4.23) is schematically represented in figure 4.13b: the solid lines are the Landau levels with index n=0, 1 and 2, that fall below the (zero field) electrochemical potential, and so which will be occupied by electrons. From the picture it is clear that in the center of the channel the constant $U(x)$ (which we take as the zero energy reference) will also make the Landau levels to be position independent. In contrast on approaching the boundaries of the system we see the electrostatic potential energy to rise and so, following eq. (4.23), also the energies of the Landau levels will increase similarly. This implies that the Landau levels cross through the electrochemical potential, one after the other and each at a distinct x-position. Before discussing the consequences of this crossing in more details let us first concentrate on the characteristic behaviour of electrons at different positions in space. An electron at the lowest (n=0) Landau level sitting in the middle of the channel (see S_1 in figure 4.13b) will see a constant electrostatic potential energy and so it effectively will behave as if it is situated in a system of infinite size. As discussed in section 4.3 this implies that it will just be confined to a disc-shaped area, as shown in figure 4.10a. Note that its center of motion will be fixed in space. Remember from section 4.3 that the diameter of this disc is approximately twice the magnetic length (eq. 4.20). The same thing holds for the higher lying state S_2 shown in figure 4.13b.

For the state S_3 at the edge (figure 4.13b) the situation is different. Here the electrostatic potential energy rises (gradually) while approaching the boundary. The gradient of this potential gives rise to a finite electric field, following $\vec{E}(x) = -\vec{\nabla}U(x)$;(note the distinction between energy E and electric field $\vec{E}$!). This has an important consequence for the motion of the electron. Looking to



a.



b.

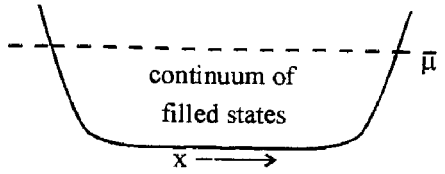
Figure 4.12. A finite width channel in a 2D electron system. In a. the channel is shown in real space; electrons are confined in the x-direction and free to move in the y-direction. Outside the channel the electron density is zero. In b. a cross section of the energy landscape in the (bounded) x-direction is shown. The dashed line is the electrochemical potential μ, which is constant throughout the system. Note the x-dependence of the electron density and so the Fermi energy.

within which conductance takes place and so the electron density is non-zero. This immediately implies that the Fermi energy will be non-zero too. As outside the channel the electron density is zero we arrive at a (x-) dependence of the electron density and the Fermi energy. This is shown in figure 4.12b. It is important to note that, in order to obtain a position-dependent Fermi energy we are not allowed to vary the electrochemical potential. In doing so we would violate the thermodynamic requirement of a constant μ throughout a system of particles that are free to move, which is the case for the electrons inside the channel. As we know that in this area the relation $\mu=E_F(x)+U(x)$ should hold, - i.e. the electrochemical potential is the sum of the kinetic energy and the potential energy - , the channel is actually formed by the position-dependence of the electrostatic potential energy, $U(x)$.

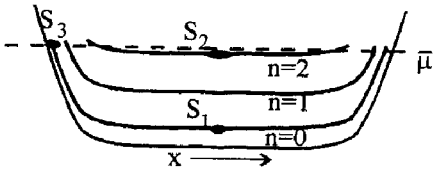
At zero magnetic field the electrons will fill all the states available up to the electrochemical potential (figure 4.13a) (Note: to complicate matters not further we assumed that the channel is so wide that so many transverse modes (see discussion in section 4.2) are formed that actually a continuum of states results). Now the magnetic field is switched on and Landau levels are formed at (kinetic) energies given by eq. (4.15) (neglecting the spin for the moment). For the total energy of an individual state in a particular Landau level it is not difficult to see that we can write

$$E_n(x,k_y)=(n+\frac{1}{2})\hbar\omega_c+U(x)+\frac{\hbar^2k_y^2}{2m} \qquad (4.23)$$

For the time being we assume the last term, representing the kinetic energy in the y-direction which is somtimes called the guiding energy, to be negligible, an assumption we will return to below. Also below it will become clear why we do not include a kinetic term for the x-direction.



a.



b.

Figure 4.13. Electron states in a channel. a. represents the case for zero magnetic field, where states are formed characterised by the wave vectors $\{k_x,k_y\}$. In b. the magnetic field is switched on and Landau levels are formed, characterised by the level indices n=0, 1, 2, ... Note the three different electron states S_1 at the lowest (n=0) Landau level in the center of the channel, state S_2 at the same position but at the highest occupied Landau level (n=2), and state S_3 in the lowest Landau level but at one of the edges.

Eq. (4.23) is schematically represented in figure 4.13b: the solid lines are the Landau levels with index n=0, 1 and 2, that fall below the (zero field) electrochemical potential, and so which will be occupied by electrons. From the picture it is clear that in the center of the channel the constant $U(x)$ (which we take as the zero energy reference) will also make the Landau levels to be position independent. In contrast on approaching the boundaries of the system we see the electrostatic potential energy to rise and so, following eq. (4.23), also the energies of the Landau levels will increase similarly. This implies that the Landau levels cross through the electrochemical potential, one after the other and each at a distinct x-position. Before discussing the consequences of this crossing in more details let us first concentrate on the characteristic behaviour of electrons at different positions in space. An electron at the lowest (n=0) Landau level sitting in the middle of the channel (see S_1 in figure 4.13b) will see a constant electrostatic potential energy and so it effectively will behave as if it is situated in a system of infinite size. As discussed in section 4.3 this implies that it will just be confined to a disc-shaped area, as shown in figure 4.10a. Note that its center of motion will be fixed in space. Remember from section 4.3 that the diameter of this disc is approximately twice the magnetic length (eq. 4.20). The same thing holds for the higher lying state S_2 shown in figure 4.13b.

For the state S_3 at the edge (figure 4.13b) the situation is different. Here the electrostatic potential energy rises (gradually) while approaching the boundary. The gradient of this potential gives rise to a finite electric field, following $\vec{E}(x) = -\vec{\nabla}U(x)$;(note the distinction between energy E and electric field $\vec{E}$!). This has an important consequence for the motion of the electron. Looking to

the problem in a classical way the combined action of the electric and the magnetic field lead to a finite *drift velocity* given by

$$v_{d,n,y}(x) = -\frac{|\vec{E}|}{|\vec{B}|} \qquad (4.24)$$

Here we take the electric field and the magnetic field perpendicular to each other, as the magnetic field is perpendicular to the 2DEG and so it will be perpendicular to any wall in the plane of the 2DEG. From the geometry it thus is clear that the (drift) velocity is directed along the y-axis, and so in the direction of the channel. So we have to conclude from this classical argument that an electron at the edge of the system will move along this edge, i.e. an *edge current* is seen to flow. It will not be difficult to see that this edge current is the quantum analagon of the current resulting from the classical skipping orbits as discussed before (see section 2.2).
Note that the drift velocity of the electron yields a wave vector in the y-direction as well, simply given by $k_{n,y} = m v_{d,n,y} / \hbar$.

Problem: Assuming a typical electric field at the edge of ~ 10^4 V/m and a magnetic field B=2 T, show that indeed the last term of eq. (4.23) is much smaller than the first one. Take the effective mass m=0.05 * the bare electron mass.

From figure 4.13 it is obvious that the direction of the velocity (and the associated wave vector) will be opposite at the two sides of the channel; this follows immediately from the opposing direction of the local electric fields. Consequently edge currents flow in opposite directions at the two sides of the channel. Note that this implies that, if the system would have been circular (or of any other shape that is finite in size) a circulating current is set up, in accordance with the simple skipping orbit picture.
Now we want to return to the discrete nature of the Landau levels and see how these edge currents affect the transport through our channel, i.e. what effect it has on the *net current* flowing from one contact at one end of the channel to a second one at the other end. From figure 4.13 we see that the state denoted by S_3 in the lowest (n=0) Landau level sits at the edge and so it will

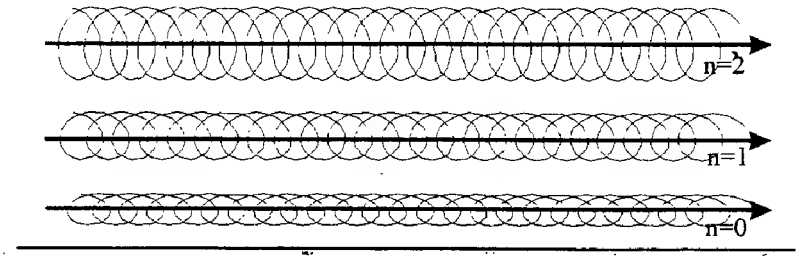


Figure 4.14. A quasi-classical presentation of edge states resulting from the orbiting+translating electron trajectories at the boundary. Based on the case of figure 4.13b the 3 filled Landau levels result in 3 distinct paths the centers of which closely follow the boundary of the system. These electron states are of a 1-dimensional nature and extend all the way along the edge, and are called edge states.

experience the lateral movement along the edge into the y-direction. Now we know that for the *net* current *only electrons at the electrochemical potential* contribute. So, we can limit our discussion to those positions in space where a Landau level crosses through the electrochemical potential, i.e. states such as S_3. Figure 4.14 shows *highly schematically* how we may view semi-classically the resulting edge state currents to flow along a (single) boundary, assuming the magnetic field and electron density of figure 4.13, i.e. with 3 Landau levels below the electrochemical potential. Note that these states extend all the way along the edge of the system. There is however one more important characteristic that has to be mentioned. From the figure we see that the typical width of the state is approximately given by the magnetic length l_B. In addition, from the discussion in section 4.3 we know that no two individual states of the same Landau level can be closer to one another than this characteristic length scale, because of the Pauli requirement. This implies that only one extended edge state electron trajectory is allowed per (crossing of a)

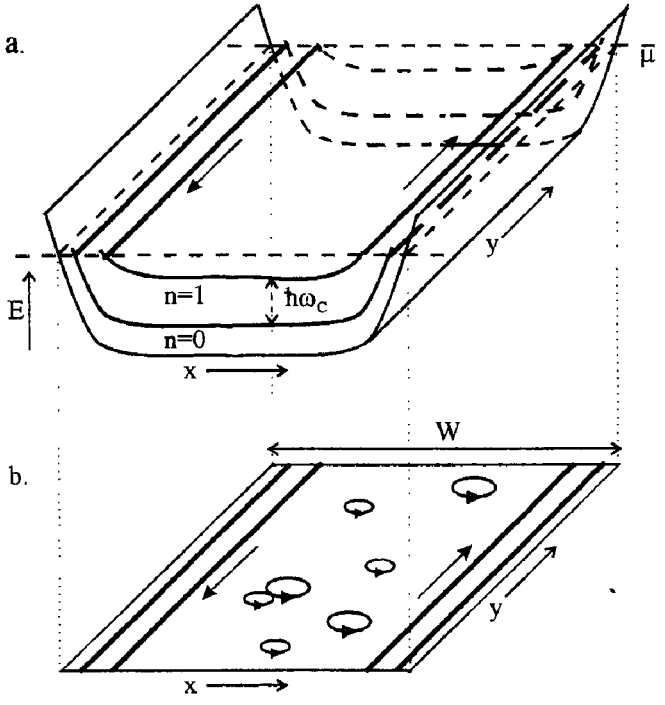


Figure 4.15. A pictorial view of edge states or edge channels in a 2DEG subject to a large magnetic field. The topmost figure (a.) shows the system in an (x,y,E) representation, demonstrating the relation between the positions of the edge states and the crossings of the Landau levels with the constant electrochemical potential. The lower panel (b.) shows these states in real (x,y,z) space. Note that the figures are cut-throughs: the system is assumed continuous along the y-direction.

Landau level, or, each edge state is *1-dimensional* in nature. This is a very important conclusion as we will see below.
So, in short, we have shown that near the boundary of a 2DEG system subject to such a large magnetic field that Landau levels are formed, well defined edge states (sometimes also called: *edge channels*) develop which are characterised by the following: they are of a 1D nature; these states have a width of typically the magnetic length; there are just as many edge states as there are filled Landau levels in the bulk (or central area) of the 2DEG; each 1D edge state is characterised by the quantum numbers $\{n,k_y\}$; they are spatially separated by distances governed by the confining potential gradient.

Problem: Evaluate the spatial separation between two adjacent edge channels, in terms of $U(x)$ and the applied magnetic field B. Taking the same typical electric field at the edge as in the preceding problem, how does the separation compare to the magnetic length?

Figure 4.15 shows in a pictorial way the most important characteristics of edge channels, both in energy space (a.) as well as in real space (b.). It illustrates the relation between the edge states and the Landau levels from which these are formed. This picture also shows that electrons residing in (Landau level) states in the center of the system, with their center-of-orbit being stationary, will not make a contribution to the currents in the system.

Now we are in the position to discuss how the fact that the net current in such a 2DEG in a large field is carried by the 1D edge states affects the conductance through the system. Figure 4.16a shows these edge states running through the channel, connecting the two contacts or reservoirs 1 and 2 with their respective electrochemical potential μ_1 and μ_2; we assume $\mu_1>\mu_2$ and write $\Delta\mu=\mu_1-\mu_2$. The left reservoir 1 (put at the largest μ_1) emits electrons into the system. Evidently

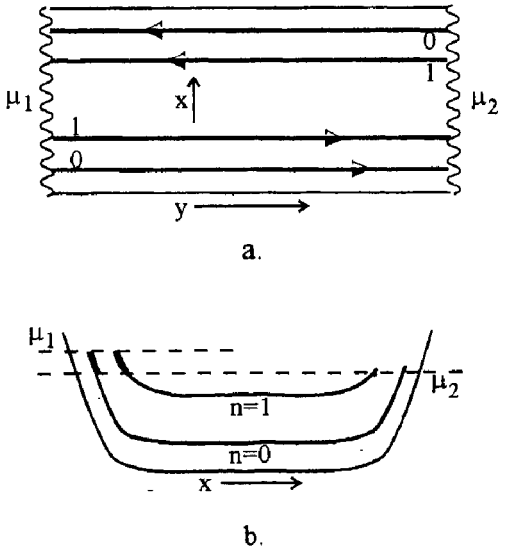


Figure 4.16. Edge states and their occupation by two contacts or reservoirs attached at the ends of the channel. Panel a. shows the system in a schematic way, with the two contacts at the ends; the contacts are taken to be at electrochemical potentials μ_1 and μ_2 respectively, with $\Delta\mu=\mu_1-\mu_2>0$. In b. the difference in occupation of the states at one boundary relative to the other is shown. Note that all the net current is carried by the edge states and in the energy interval $\Delta\mu$.

these electrons only can occupy edge states. Now it is important to note that at only one side of the confining channel these electron states allow electrons to travel from contact 1 to contact 2. So, only at *one side of the channel* the edge states will become occupied up to the electrochemical potential μ_1 of the emitting contact 1. By the same token the edge states at the opposite side of the channel only contain electrons travelling in the opposite direction and they will be occupied by electrons emitted from contact 2 and thus these states will be filled up to μ_2. Note that again the magnetic field leads to breaking of the time-reversal symmetry: while at zero field all states can be occupied by electrons travelling in both directions this is no longer so at non-zero magnetic fields. Figure 4.16b shows this in an (E,x) cross section through the 2DEG. For simplicity two Landau levels are assumed to be occupied in the center of the 2DEG, and so two edge states will be present.

To evaluate the current we simply can take the same approach as before in section 4.2, starting from eq. (4.2). For each edge channel n we can simply write

$$J_{net,n} = -e \int_{\mu_2}^{\mu_1} v_n(E) * \rho(E) dE \qquad (4.25)$$

for the net current as it is transported within the energy interval $\Delta\mu=\mu_1-\mu_2$. Note that, as before, all the current from e.g. contact 1 to 2 for energies $<\mu_2$ is compensated by current (at the opposite edge) running from contact 2 to 1, and so this will not contribute to the net electron flow from contact 1 to 2 (and so the net current from contact 2 to 1 (because of the negative electron charge)). Now the important aspect of the 1D nature of the edge states comes into play. From the discussion around eqs. (4.4) and (4.5) we know that the particular feature of 1D systems is that the velocity and density-of-states have their energy dependence cancel, yielding that the product of the two is the constant $1/h$ (see eq. (4.5)). This is also the case here, and so we can use all the steps following eqs. (4.4) and (4.5) as well. This immediately leads to the net current *per edge channel* being given by

$$J_{net,n} = -eg_s \frac{1}{h} \Delta\mu = g_s \frac{e^2}{h} V_{bias} \qquad (4.26)$$

with $V_{bias}=-(\mu_1-\mu_2)/e$ being the voltage difference between the two contacts 1 and 2. This is the current per edge channel, and so the total current is obtained by simply multiplying (4.26) by the number of edge channels (or occupied Landau levels), N_{max}. Consequently the (two terminal) conductance is given by

$$G = \frac{J_{net}}{V_{bias}} = \frac{\sum_{n=0}^{N_{max}-1} J_{net,n}}{V_{bias}} = N_{max} \frac{g_s e^2}{h} \qquad (4.27)$$

which again shows the (by now rather familiar) quantisation of the conductance in units e^2/h (if the spin is taken into account). The "quantisation counter" N_{max} is now given by the number of edge states in the system, in contrast to the zero magnetic field case discussed in section 4.2 where this parameter was related to the width W of the system (see eq. (4.11)).

As a next step we introduce additional contacts in the system. More specifically we add two more contacts in order to realise a system in which we can study the Hall effect. Figure 4.17 shows the typical Hall configuration, with a total of 4 contacts (note the similarity to e.g. fig 3.11). On

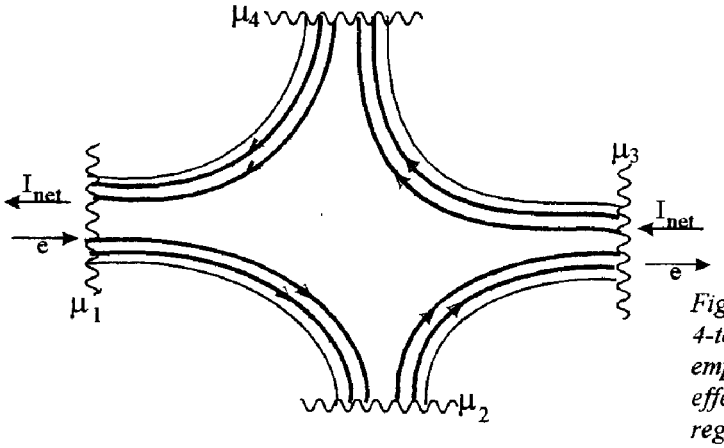


Figure 4.17. A typical 4-terminal configuration employed to discuss the Hall effect in the edge channel regime. For simplicity two edge channels are assumed to exist (N=2).

purpose we have deformed the ideal Hall cross into a topology that is less symmetric, as we want to show that the result we obtain for the Hall conductance will be extremely insensitive to the particular configuration. This is in marked contrast to a common Hall experiment, where the symmetry of the layout is crucial for a valuable measurement. We assume the net current to be introduced via the (current) contacts 1 and 3, and we evaluate the (Hall) voltage developing between the (voltage) terminals 2 and 4. Take $\mu_1 > \mu_3$.

As *all* edge states populated by reservoir 1 up to μ_1 follow the lower edge and terminate in (voltage) contact 2 the *total* current running from 1 to 2 is given by

$$J_{1\to 2} = \mu_1 g_s \frac{e}{h} N \qquad (4.28a)$$

It is crucial to note that the *direct* inverse current, i.e. flowing directly from reservoir 2 to reservoir 1, is *zero*, as no edge state leaving reservoir 2 connect to contact 1. Or

$$J_{2\to 1} = 0 \qquad (4.28b)$$

Similarly for the total currents injected into the system from all contacts we can write down the same, or

$$J_{i\to i+1} = \mu_i g_s \frac{e}{h} N \qquad (4.29a)$$

and

$$J_{i+1\to i} = 0 \qquad (4.29b)$$

Now we have to distinguish between the current contacts 1 and 3 and the voltage contacts 2 and 4. At voltage contact 2 the *net* current should be zero and so

$$J_{1\to 2} = J_{2\to 3} \qquad (4.30a)$$

or from eq. (4.28a)

$$\mu_2 = \mu_1 \qquad (4.30b)$$

Evidently the same will hold for contacts 3 and 4 and so

$$\mu_4 = \mu_3 \qquad (4.31)$$

Now let us turn to the current contacts. Here the conservation of current requires (see figure 4.17)

$$I_{net} = -(J_{1\to 2} - J_{4\to 1}) \tag{4.32}$$

and so, using eqs. (4.28a) and (4.30b) we immediately find

$$I_{net} = -(\mu_1 - \mu_3) g_s \frac{e}{h} N = -(\mu_2 - \mu_4) g_s \frac{e}{h} N \equiv e V_{24} g_s \frac{e}{h} N \tag{4.33}$$

So, the Hall conductance follows as

$$G_{Hall} \equiv \frac{I_{net}}{V_{24}} = N g_s \frac{e^2}{h} \tag{4.34}$$

We find that the Hall conductance (or equivalently the Hall voltage at fixed current) is quantised, again the well-known units e^2/h. This is the famous *integer quantum Hall effect* or *IQHE*, as it was found experimentally by K. von Klitzing et.al. in 1980 (Phys. Rev. Lett. 45, 494 (1980)), an accomplishment for which von Klitzing received the Nobel prize for physics. It is denoted "integer" while the counting factor preceding the quantised unit e^2/h is an integer, a point we will return to at the end of this section.

One of the more recent experimental results of this quantisation of the Hall conductance is shown in figure 4.18. Even though this figure may look somewhat complicated at first sight it contains a wealth of information which we will discuss step by step. Two traces are displayed, the topmost one being the Hall conductance $G_{\rm Hall}$, while the lower, much more structured trace denoted by ρ_{xx} basically is a measure for the strength of the scattering from edge to edge (i.e. inter-edge events of the type S_2 or S_3 that will be discussed in figure 4.19). The filling factors are indicated at the top of the figure. Note that the *plateaus* in the $G_{\rm Hall}$-B curve correspond to the complete filling of a certain number of Landau levels (i.e. *integer filling factors* ν, as given by eq. (4.18)) while the steps in between correspond to the case that a certain Landau level in the bulk (or center of the channel) passes through the electrochemical potential. Note also that the structure in both curves does not stop at filling factor ν=1, but in stead a whole "wood" of peaks and valleys, and plateaus is visible at filling factors smaller than one: we will return to these features at the end of the chapter.

In order to appreciate the quite striking implications of this (integer) quantisation we need to discuss a few more details. First, it is of importance to note that in the derivation along the eqs. (4.28) to (4.34), we have actually employed the Landauer formalism as introduced in section 4.2 around eq. (4.12). The particularly attractive feature we were able to use was that the transmission probabilities T_{jk} were very easy to evaluate, as these were either 1 or 0 (e.g., $J_{1\to 2} \neq 0 \Rightarrow T_{1\to 2} \equiv T_{21} = 1$ and $J_{2\to 1} = 0 \Rightarrow T_{2\to 1} \equiv T_{12} = 0$).

The second point to note is that at first sight it may seem rather "superfluous" to make the step in eq. (4.33) from μ_1 to μ_2 and from μ_3 to μ_4. This however is not the case. It should be note that the determination of the electrochemical potential of a *current carrying* reservoir is not at all obvious: any non-zero resistances inside the contact will lead to potential gradients and so it is not clear what will be the *exact value* of the electrochemical potential up to which energy the edge states will be occupied. This is no longer so for the voltage contacts. Here the zero current condition allows a direct measurement of the electrochemical potential of the *emitted* electrons (i.e. those that are injected into the system). It is this distinctive use of current and voltage contacts which allowed the first quantised Hall measurement by von Klitzing et.al. already to be at a precision of ~ 0.001 % or 10 part-per-million (ppm). At present (1996) such measurements are employed to realise a universal standard of resistance, and accuracies of 0.001 ppm (or nine digits behind the decimal point!) have been achieved.

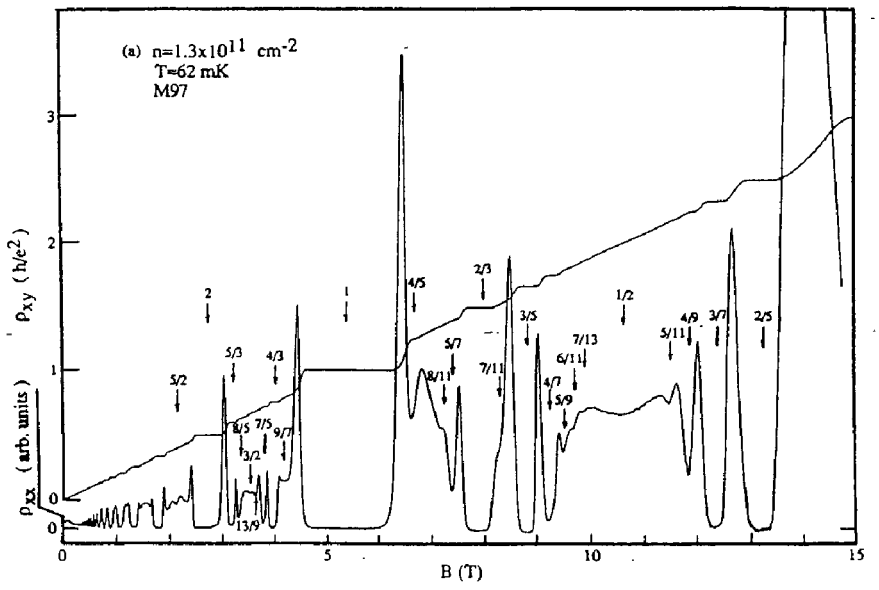


*Figure 4.18. The quantised Hall effect in 2DEG in a GaAs-AlGaAs heterostructure. The very high quality (μ=200 m^2/Vs or l_e~10 μm) of the low-density electron system (n_e=1.3*10^{15} m^{-2}) is clearly seen from the rich structure in both the Hall resistance ρ_{xy} as well as the length-wise resistance ρ_{xx}. In the Hall trace flat plateaus are seen: these occur at integer filling factors ν, and demonstrate the quantum Hall effect. Note that a plateau in the Hall resistance corresponds to a minimum in ρ_{xx}. Based on the electron density ν=1 is found at B=5.5T. Up to this field value the plateaus result from the integer quantum Hall effect; beyond this field the plateas (and corresponding minima in the resistance) are due to the fractional quantum Hall effect. The feature at B=11T at ν=1/2 is discussed separately in the text.*

In this respect a third comment is useful. For this accuracy to be reached one has to have unity transmission from contact to contact. We want to dwell on this aspect somewhat more, as it will provide us with some more insight in the effect of scattering on the QHE. Figure 4.19 shows the Hall cross configuration with again two edge channels. In this figure a number of different scattering processes are indicated and we will see what their specific consequences are for the QHE. One can distinguish between intra-edge scattering events, making the electron to change from one edge state to a different one (i.e. a change of *n*) however continuing to adhere to *the same boundary* of the system; one such case is indicated by S_1 in figure 4.19. It is crucial to note that this would only lead to a rearrangement of the charge carriers over the different edge states, and so it will *not* affect the total current that enters the following (voltage) contact. This implies that intra-edge scatterings do not affect the accuracy of the QHE. The same holds for scatterings such as S_2 and S_3. These events, where the particle crosses all over the sample from one boundary to the opposite one and called inter-edge scattering, indeed affect the net current. However the current arriving at the voltage contact is affected by *exactly the same amount* and so the Hall conductance will remain unaffected. The third type of events, denoted by S_4, however is of a more serious nature . This is also an inter-edge event, however now the net current is *not* affected while

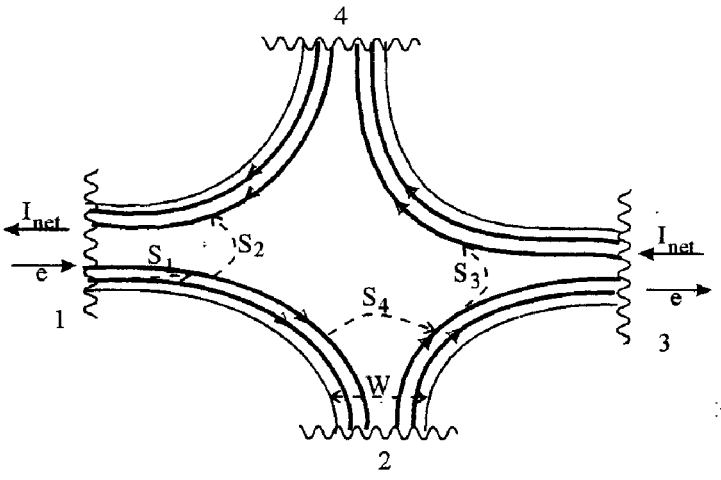


Figure 4.19. Scattering processes in the quantum Hall regime. Note the distinction between intra-edge scattering events (such as S_1) and inter-edge scatterings (S_2, S_3 and S_4). In particular events that allow an electron to by-pass a voltage contact strongly affect the QH accuracy, e.g. process S_4 which pass by contact 2.

the current arriving at the voltage contact (in this case contact 2) will become *reduced*. Thus the Hall conductance will increase beyond its quantised value, detoriating the intrinsic QH accuracy. It is good to note in returning to figure 4.18 that indeed we see that at the plateaus of the Hall conductance, where the filling factor is integer, the value of ρ_{xx} becomes very small. This is in agreement with the fact that ρ_{xx} represents the degree of inter-edge scattering and the notion that for an accurate quantisation the inter-edge scattering should be minimal.

Problem: what geometrical measures should be taken in order to reduce inter-edge scattering processes as much as possible? What length scale(s) are involved?

4.5. Closing remarks

Before closing this chapter we want to hint upon a few more phenomena that relate to the integer quantum Hall effect. It is impossible to discuss these at any depth as most of these phenomena require a much more rigid approach. The essential difference between the simple integer effect and individuals from this "zoo of exotic (though very interesting!) animals" requires one to go beyond our beloved single-particle picture for their proper description.

First, the very fact that we indicate the phenomenon discussed up to now as the *integer* QHE indicates that a different type may exists as well. Indeed, only a few years after the 1980 discovery of the *I*QHE, a second type was found (D.C. Tsui et.al, Phys. Rev. Lett. 48, 1559 (1982)). In large magnetic fields such that the filling factor $\nu = n_e h/eB$, representing the number of electrons per flux quantum (see eq. (4.18)), is <1, additional flat plateaus are found. Figure 4.18 shows a whole sequence of these to occur. From analyses it is found that these corresponds to filling factors of a *fractional* value

$$\nu = \frac{p}{q} \qquad (4.35)$$

with p and q positive integers, and q being restricted to *odd values*. As the filling factors no longer are integer, this effect is denoted as the *fractional quantum Hall effect* or *FQHE*. The essential difference with the integer effect is that in the description of the fractional effect the Coulomb interaction of the electrons has to be taken into account. Note again that at the plateaus in the Hall conductance the scattering is suppressed, as is seen from the strong reduction of ρ_{xx}.

The second, even more recent addition is only slightly visible in figure 4.18, albeit only in the ρ_{xx} trace. At such a magnetic field that the filling factor equals 1/2 a weak feature is seen. More importantly, looking to the symmetry of the FQH ρ_{xx}- minima structure below and above the ν=1/2 position makes one to believe that, even though the 1/2 state seems to differ from the regular FQH states (e.g., as seen from the absence of a plateau at 1/2), it at least will not be independent from it. This notion has been addressed strongly over the last few years and a rather peculiar picture has emerged, based both on theoretical considerations as well as experimental results. It is believed that at (and near) filling factor ν=1/2 a new particle comes into being. This particle is a combination of an *electron with two flux quanta coupled* to it. Such a dressed-up entity is denoted a *quasi-particle*, and for this particular case it is called a *Composite Fermion* or *CF*. These peculiar particles have been proposed first by J.K. Jain in 1989.
Now, as this CF particle has two flux quanta adhered to it, at a field such that ν=1/2 the total available field is "distributed" over the available electrons (to form the CF's), and as a consequence the *CF experiences a zero effective magnetic field.* One of the immediate consequences is that at such a field that ν=1/2 the CF should travel in a *straight line*, or the composite particle will show an infinitely large cyclotron radius. In section 3.3a the electron focusing experiment was performed at small fields to verify the classical cyclotron motion of electrons. A similar experiment should be possible with composite Fermions, but now with reference to *effective* fields. Indeed such CF focusing experiments have been done to investigate this prediction and has been found to agree.
Just as a short final note on these peculiar particles we return to the noted particular grouping of the fractional ρ_{xx}- minima around the 1/2 filling factor field value. These fractional states now can be interpreted as the *integer quantum Hall states* for the *Composite Fermion*. To show this more clearly one can rewrite the fractional filling factors (4.35) for the most clearly visible plateaus (or the strongest minima in ρ_{xx}) as

$$\nu = \frac{p}{2p \pm 1} \tag{4.36a}$$

or, writing ν_0=1/2 (i.e. the filling factor for the ideal formation of the CF's)

$$\frac{1}{\nu} = \frac{1}{\nu_0} \mp \frac{1}{p} = \frac{p/\nu_0 \pm 1}{p} \equiv \frac{2p \pm 1}{p} \tag{4.36b}$$

This expression shows that p now represents the index of the highest filled Landau level, however not for electrons but for our new quasi particle, the CF.
It will not come as a surprise that in this highly intriguing much work is in progress to understand the properties of these exotic new particles.

From the preceding two examples it will be clear that exciting new phenomena are still discovered in these by now very mature 2-dimensional electron gases in semiconductors, and it is likely that more will emerge in the (near) future!

General references to the subject

1. "Quantum Transport in Semiconductor Nanostructures" by C.W.J. Beenakker and H. van Houten, in Solid State Physics, ed. H. Ehrenreich & D. Turnbull, Vol 44 (Academic Press Inc., Boston, 1991), sections 4, 12 and 18

2. "Electric Transport in Mesoscopic Systems", S. Datta (Cambridge University Press, Cambridge, 1995), chapter 4

3. "Half-filled Landau level yields intriguing data and theory", B. Schwarzschild, Physics Today (July 1993), p 17-20; "How real are Composite Fermions?", W. Kang et.al., Phys. Rev. Lett. 23, 3850 (1993)

5. Aharonov-Bohm effect and persistent currents in non-superconducting systems

Keywords: phase coherence; fluxquantum; diffusive transport; thermodynamic equilibrium; energy eigenstates;

5.1 Introduction

In the classical Drude description of transport of electrons in the solid state (see chapter 2, section 2) it is assumed that, whenever an electron scatters, all information on its previous state will be lost. As we have seen in chapter 1 such loss of information only occurs in the case of inelastic scattering events, where each particular scattering event is unpredictable in both time and position. In contrast, elastic scattering affects the trajectory of the travelling electron in such a way that the phase shift resulting after the scattering event can be predicted. Consequently this will lead to quantum interference of the electronic wavefunction over a length scale determined by inelastic processes, a size which commonly is much larger than the elastic mean free path (e.g., see chapter 2, section 3). In addition, this not only holds for a single electron, but also for the average effect of a number of electrons, i.e. it affects the conductance.
In this chapter we will discuss two particular effects which clearly demonstrate this quantum interference, both occurring in small ring-shaped structures. These are the Aharonov-Bohm effect and the phenomenon of persistent currents in normal (i.e. non-superconducting) conductors. The two effects are discussed here sequentially as they are connected rather strongly.

5.2 Aharonov Bohm effect in the solid state

In this section we introduce the effect of a magnetic field on the electron transport through ring-shaped conductors at low temperatures, such that phase coherence of the electron wavefunction in the ring is (at least partially) preserved. We will see that the flux enclosed by the ring modulates the total transmission through the ring with a period of either one flux quantum or half of it. This periodic flux dependent modulation leads to an oscillatory behaviour of the conductance versus the applied B-field. The amplitude of this oscillatory part is found to be $\sim 2e^2/h$ at low temperature and low voltage across the ring terminals. Increasing these leads to suppression of this amplitude, and we will see that the characteristic energy scale is given by the Thouless energy, introduced in chapter 1. We will also see how averaging over an ensemble of (nominally equal) rings will lead to suppression of the one-flux quantum periodicity while preserving the half-flux quantum contribution. After a brief introduction to the effect, we continue with a theoretical discussion of the subject in subsection 5.2.a., followed by a few important experimental results (subsection 5.2.b.).

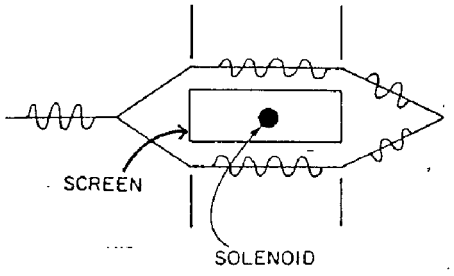


Fig. 5.1. Aharonov-Bohm configuration. The solenoid generates a magnetic field, with the shield preventing the field to be present at the trajectories of the electron.

The Aharonov-Bohm (AB) effect bears its name after Aharonov and Bohm who presented a thought experiment, describing how a magnetic flux affects the interference of a split electron wave [Phys Rev. 115, 485 (1959)]. The most essential aspect of the "experiment" is that it unambiguously confirms the physical reality of the *magnetic vector potential*, raising it from a mathematical entity to a true physical quantity. Fig. 5.1 shows the basic features of the experiment. An electron wave emitted from a source is split into two partial waves which are recombined at a further distance, effectively forming an electron interferometer. The two partial waves enclose a flux localised in the area between the two waves, such that *no magnetic field exists anywhere along the path of the electron waves*. However, despite the absence of a magnetic field all the way along the trajectories (and so *without* the Lorentzforce acting on the particle!), the *vector potential* associated with the field can not taken to be zero everywhere along a closed path, irrespective what gauge is chosen.
Quantum mechanically we know that the vectorpotential A modifies the canonical momentum of an electron with charge $-e$ travelling at a velocity v according to

$$p = hk/2\pi = mv + eA \qquad (5.1)$$

A change in momentum implies a change in phase acquired by the electron after travelling a certain distance (see also below). So, a magnetic field which may be zero at the actual electron trajectory, still affects the phase of the electron wave via the vector potential.
It took quite some time before this rather remarkable fact became accepted, and even now the discussion is not fully completed, despite the indisputable experimental confirmation of the effect for electrons in free space (A. Tonomura et.al., Phys Rev. Lett. 48, 1443 (1982); A. Tonomura et.al., Phys. Rev. Lett. 56, 792 (1986)). Although of considerable fundamental interest, we will not discuss these experiments in detail but continue with the theory for electrons in a solid.

The total phase acquired by a propagating electron along some path described by a position variable l immediately follows from the canonical momentum of eq. (5.1)

$$\Delta\phi = \int \vec{k}.d\vec{l} = \frac{1}{\hbar}\int (m\vec{v} + e\vec{A})d\vec{l} = \Delta\phi_v + \Delta\phi_A \qquad (5.2)$$

i.e. the sum of a contribution due to the velocity v of the particle and one resulting from the vectorpotential A. Thus, the phase shift induced by the vectorpotential is given by

$$\Delta\phi_A = \frac{e}{\hbar}\int \vec{A}.d\vec{l} \qquad (5.3a)$$

and so the phase shift induced by the vectorpotential on an electron propagating once around a closed loop is found as

$$\Delta\phi_A = \frac{e}{\hbar}\oint \vec{A}.d\vec{l} = \frac{e}{\hbar}\iint (rot\vec{A}.d\vec{S} = \frac{e}{\hbar}BS = 2\pi\frac{\Phi}{\Phi_0} \qquad (5.3b)$$

with Φ_0 being the flux quantum h/e. So the phase difference equals 2π *times the enclosed flux in units of the flux quantum* Φ_0. As a phase difference is only distinguishable Mod(2π) *any* effect resulting from this enclosed flux will show a periodic behaviour, with a period of one flux quantum. The enclosed flux is often referred to as the *Aharonov-Bohm-* or *AB-flux.*

5.2.a. *The Aharonov-Bohm and Altshuler-Aronov-Spivak effects in the solid state*

We want to investigate how the Aharonov-Bohm effect manifests itself in the solid state. We concentrate on electrons in metallic structures in the quantum diffusive regime, i.e. $l_e << l_\phi \sim < L$, with L denoting the size of the AB structure, more specifically the length of the circumference of the ring. Contrary to the original AB experiment, in these structures the magnetic field commonly is not confined to the centre of the ring or loop only, but it will also penetrate the metal where the electrons reside, i.e. the electrons experience the B-field in this case. Despite this (rather fundamental!) difference, also in the solid state the effect is called the Aharonov-Bohm effect.
Figure 5.2 shows a ring of small crossection, with two diametrically positioned leads attached at the junctions J_1 and J_2 to allow conductance measurements. An external magnetic field induces a flux Φ in the ring. Two types of trajectories are shown, one connecting the input junction (J_1) to the output (J_2), and the second type returning the particle to the input junction. In figure 5.2a two paths 1 and 2 of lengths L_1 and L_2 respectively ($L_1 \sim L_2$) run along the "top" and "bottom" branch of the ring, each path covering only half of the circumference L of the ring. In b. the paths 3 and 4 fully encircle the loop clockwise and counter clockwise respectively.
Let us first concentrate on the trajectories of the type 1 and 2 shown in figure 5.2a. The acquired difference in AB phase between the two paths follows directly from eq. (5.3b), with half of the total AB phase lagging (i.e. with negative sign) along one path and the other half leading (positive sign) along the second path. Consequently the phase difference will increase

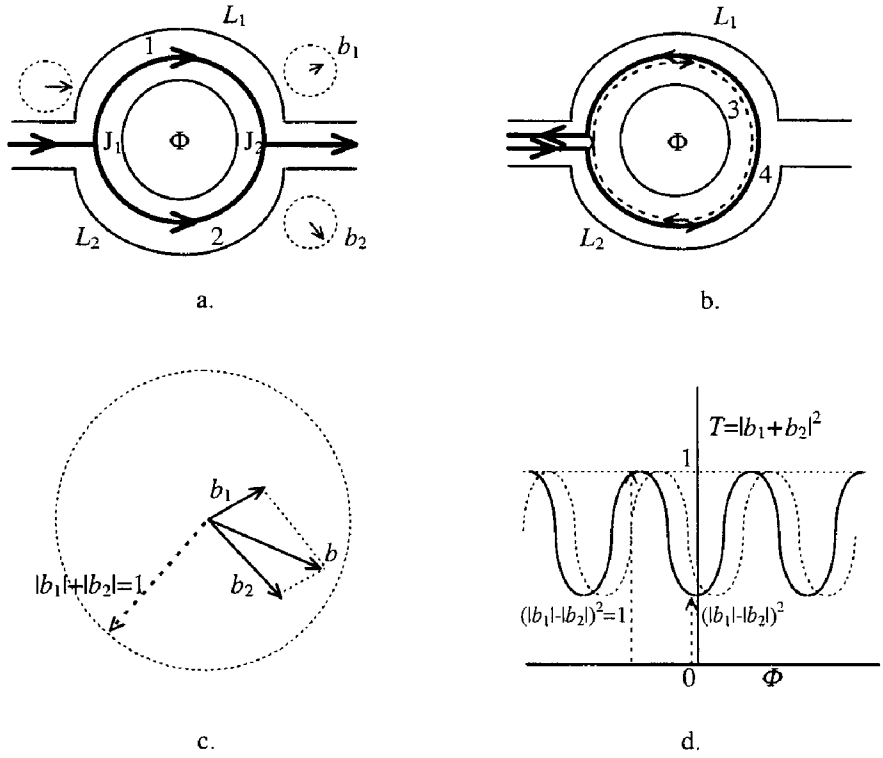


Figure 5.2. Simplified picture of the electron wave interference in a conducting ring. Two types of trajectories contributing to the AB effect are shown. J_1 is (arbitrarily) taken as the junction with the input lead and J_2 that for the output side; a.: two semicircular orbits 1 & 2 leading to coherent forward scattering; b.: two full circular orbits 3 (dashed line) & 4 (solid line) leading to coherent backscattering; c. vectorial summation of output wave amplitudes; d. total transmission probability $T=|b_1+b_2|^2$ (solid line is for identical path lengths 1 & 2; dashed line for non-identical path lengths). A more elaborate evaluation shows that the oscillation at $\Phi=0$ can not take any random value between the extrema, but is either maximum or minimum, i.e. the phase is either $\Delta\phi=0$ or $\Delta\phi=\pi$. This is shown in e.

by 2π each time one flux quantum is added to the ring. So the period of the AB oscillations resulting from these types of trajectories equals Φ_0 or h/e.

In chapter 2 we have seen that, based on the Landauer approach, the conductance of a system is determined by the transmission probability of the electron waves. So, to calculate the conductance of the ring we have to evaluate the total transmission probability for a unity incoming wave from (e.g.) the left lead, splitting at the entrance junction J_1, each partial wave travelling along its respective path and recombining at J_2, *taking into account the phase acquired by each partial wave.* We first will give a simplified picture of the resulting interference of the electron waves, and discuss and correct one error that results from this simplified picture afterwards.
Let us write a_1 and a_2 for the amplitudes of the partial waves in path 1 and 2 leaving the wave-splitter J_1, and b_1 and b_2 for the waves arriving at J_2 to form the output wave (fig. 5.2a). With unity incoming wave, $|a_1+a_2|^2=1$ the total transmission probability T is derived from the *vector summation* of the partial waves b_1 and b_2, yielding $T=|b_1+b_2|^2$ (figure 5.2c and d). We simply can write

$$b_j = a_j F(L_j)\exp(-i\Delta\phi_j) = a_j F(L_j)\exp(-i(\Delta\phi_{v,j} + \Delta\phi_{A,j})) \qquad (j=1,2) \qquad (5.4)$$

with $\Delta\phi_{v,j}$ denoting the acquired phase at zero B field for the partial wave along path j, etc. The prefactor $F(L_j)$ is real and we assume it to be 1 for the moment.
For the transmission probability T we thus find

$$T = a_1^2 F^2(L_1) + a_2^2 F^2(L_2) + 2a_1 a_2 F(L_1)F(L_2)\cos(\Delta\phi_1 - \Delta\phi_2) \qquad (5.5)$$

In an ideal, perfectly symmetric ring both the lengths of the trajectories 1 and 2 as well as the amplitudes of the partial wave amplitudes a_j for the two paths ($a_1 = a_2 = \frac{1}{2}\sqrt{2}$) are equal. The equality of the length implies that also the phase differences $\Delta\phi_{v,j}$ for the two paths will be the same. At zero magnetic field the two partial waves will arrive at the output terminal with the same phase, i.e. $\cos(\Delta\phi_1 - \Delta\phi_2)=1$. From eq. (5.5) we see that this constructive interference yields a *maximum* total (forward) transmission probability T of the wavefunction at the output lead, or *a maximum of the conductance* through the ring (fig. 5.2d, solid curve). With the assumption $F(l)=1$ we find $T=1$, i.e. unity conductance at $B=0$ or $\Phi=0$. At non-zero flux a non-zero phase difference will result from the vector potential term and so the cosine term in eq. (5.5) will become <1 (fig. 5.2c). In particular at $\Phi=\Phi_0/2$ the cosine will be -1, leading to a full suppression of the transmission, i.e. zero conductance or infinite resistance of the loop. Thus, the resistance is found to oscillate with the applied flux, with a period given by a flux quantum penetrating the ring. The size of the oscillation is governed by the ratio of the two output partial waves b_1 and b_2.
To fulfil the conditions for a ring to be ideal we have to realise that the zero flux phase difference result from the difference in path length of the two halves of the ring, taken in units of the Fermi wavelength λ_F. In a metal λ_F equals a few tenths of a nm. This would imply that the fabrication of these rings should be accurate to better than 0.1 nm, a condition which can not be met using present day technology (see Appendix A). So, in a realistic ring the two halves of the loop will not be identical and thus the phase difference on arrival at the output at zero B-field will not be zero. As the phases will depend on the *specific microscopic details of the two trajectories*, its difference $\Delta\phi(\Phi=0)$ at arrival will have a *random value* anywhere between $-\pi$ and $+\pi$, making the total transmission to lay

somewhere in between the maximum and minimum *at random at* Φ=0 (fig. 5.2d, dashed curve).
The picture we presented above provides a transparent view of the way electron wave interference leads to AB oscillations. However, as we mentioned at the beginning, it is highly simplified. If a more elaborate approach is taken the main modification that follows relates to the zero-flux phase of the oscillation. In the preceding discussion we "forgot" to take into account that, at (e.g.) the exit splitter a certain fraction of the wave that approaches the splitter via one arm does not leave the ring but "crosses over" and continues into the other arm, somewhat similar to the AAS type of contribution. This has consequences for the zero-flux phase. While the simple picture yields a random value for it, i.e. $-\pi<\Delta\phi(\Phi=0)<-\pi$, the more complete calculation shows that $\Delta\phi(\Phi=0)$ is *either* 0 *or* π. Now it is found that the randomness associated with the microscopic details of the ring enters in a different way, namely *which* of the two values is assumed in a particular case is again *random*. This is shown in figure 5.2e by the solid and dashed curves. As just briefly mentioned, it is the accounting for the wave that continues inside the ring after a splitter that is the key element for this *symmetry* of the transmission (and conductance) around zero flux. Basically this "accounting" guarantees that we meet the fundamental requirement of conservation of particles in the transmission process through the structure, with in addition the condition for time-reversal symmetry at zero magnetic field. Note that the first of the two conditions is quantummechanically equivalent to the conservation of wavefunction probability $|\Psi|^2$. One step further, the correct way of keeping track of the wavefunction probability implies that we allow the formation of *well-defined eigenstates* in the ring. We will return to this aspect in section 5.3 on persistent currents in normal (= non-superconducting) conductors, where the existence of these eigenstates is at the key element for the occurrence of such currents.

In a realistic ring two more influences have to be accounted for. Just as the lengths of the two halves may differ, also the junctions may show asymmetries in their wave splitting (and combining) properties. This implies that the amplitudes a_j will differ. From eq. (5.5) it can be seen immediately that T_{max} will be kept equal to 1. However, the minimum transmission will become finite, i.e. $T_{min}>0$. This implies a *reduction* of the amplitude of the AB resistance (or conductance) oscillation (see figure 5.2d).

In our preceding discussion we assumed that the *total* incoming wave intensity arrives at the output junction, i.e. in eq. (5.4) we took the prefactors $F(L)$=1. This will no longer hold when *scattering* is taken into account: any scattering of the electron(-wave) while traversing the arms, either elastic or inelastic, will reduce the transmitted amplitude, making $F(L_j)<1$. However, the way the two different scattering mechanisms affect the transmitted amplitudes differs considerably.
Inelastic scatterings will reduce the phase coherence during the travelling of the waves along the respective paths. So, the resulting *phase coherent amplitude*, which is the *only relevant contribution* to the AB effect, will be governed by an exponential decay in dependence of the path length l in terms of the phase coherence length l_ϕ i.e. given by $\exp(-L_j/l_\phi)$. This loss of phase coherence thus is accounted for by multiplying the total transmission T of eq. (5.5) by a factor

$$G(l_\phi)=exp(-L_1/l_\phi).exp(-L_2/l_\phi)=exp(-L/l_\phi) \qquad (5.6)$$

with L denoting the length of the circumference of the ring. From this it is clear that the phase coherence length is the relevant length scale governing the maximum size a loop may have to display these phenomena, i.e. preferably the diameter should be not too large compared to l_ϕ.
The effect of *elastic* scattering on the amplitude is not so simple to evaluate quantitatively. Initially an elastic scatterer will reduce the amplitude of the forward travelling wave in an arm. However, the wave thus reflected may become re-reflected by a scatterer lying more "upstream" along the path. This process results in multiple elastic scattering leading to a complicated standing wave pattern, the details of which will depend on the *microscopic spatial arrangement* of the individual elastic scatterers. Evidently this process affects both the (zero B-field) phase and amplitude of the wave that ultimately arrives at the output junction, however its complexity prevents a quantitative evaluation. We will return to the problem of the amplitude of AB oscillations after discussing the related AAS oscillations and see that it relates to the problem discussed at the very end of chapter 2, when we very briefly hinted upon Universal Conductance Fluctuations (UCFs).
Despite the complexity introduced by elastic scatterings this has *no effect* on the phase*shift* induced by the magnetic field. This rather remarkable (and counterintuitive) fact is an immediate consequence of the way the vector potential acts along a path, as expressed by eq. (5.3b). From it we read that the field-induced phase shift only relies on the enclosed area S of a closed orbit. So, irrespective of the details of the trajectory (e.g. whether parts of it are traversed more than once due to multiple reflection), at a given field value the acquired phase of eq. (5.3b) will depend only on the fixed area enclosed by the circumference of the ring.

Let us now turn to the second type of trajectories, indicated by 3 and 4 in Fig 5.2b. In contrast to the preceding case these waves encircle the *full* loop, returning to their point of entrance to interfere. This leads to *coherent back scattering* (or reflection), in contrast to the preceding semi-circular trajectories showing coherent forward scattering or transmission (note that both affect the total transmission, as the sum of the probabilities for reflection and transmission equals 1). Apart from this, a more intricate difference between the two types arise.
Assuming only *elastic* scattering processes in the two arms, *any* (clockwise) path 3 can also be traversed in the reverse (counter clockwise) direction, i.e. being a path of the type 4. So, each path 4 can be seen as the *time reversed* trajectory of a path 3, making the phases acquired (at zero flux) for the two partial waves *exactly equal*. This implies that at zero flux the resulting interference will *always* yield the *maximum* total wave amplitude, or maximum *backscattering* probability R. With T=1-R this yields a minimum in the transmission probability from entrance to exit lead and so a maximum in the resistance (minimum in the conductance) of the loop.
At a non-zero flux Φ the complete encircling of the loop (in mutually opposite directions!) leads to the following phase shifts for the waves 3 and 4:

$$\Delta\phi_3 = +2\pi\frac{\Phi}{\Phi_0} \qquad (5.7a)$$

and

$$\Delta\phi_4 = -2\pi\frac{\Phi}{\Phi_0} = -\Delta\phi_3 \qquad (5.7b)$$

implying a total phase difference $\Delta\phi_3$-$\Delta\phi_4$ that is *twice as large* as in the preceding case of the semi-circle trajectories 1 and 2. So, in this case one oscillation period is obtained if the flux through the loop is increased by $\Phi_0/2$, i.e. twice as *small* as before. This is called the $\Phi_0/2=h/2e$, Altshuler-Aronov-Spivak or AAS oscillation, after the persons who were the first to investigate these.

Concerning the effect of scattering all the arguments given before for the common AB effect also apply to the present AAS case.

Problem: The effect of the phase coherence length is to reduce the AB resistance oscillation amplitude. Discuss which will be affected most: the *h/e* or the *h/2e* oscillations.

We now want to return to the problem of how to evaluate the amplitude of AB (and AAS) oscillations. In addition it is highly instructive to discuss how this depends on external parameters such as temperature and applied current or voltage (see also Appendix B). In the preceding discussion we have evaluated the conductance based on the transmission probability of *a single* electron wave (see e.g. eq. (5.6)). We thus did assume that the structure was 1D. This implies that the width of the ring should be on the order of half a Fermi wavelength, i.e. ~0.1 nm in a metal and ~10-20 nm in a 2DEG semiconductor. In real experiments the minimum width of the ring is dictated by fabrication capabilities, so ~30 nm (see Appendix A). So from this experimental constraint it is evident that the single-channel 1D description can be seen as approximate only. The effect of the finite phase coherence length is not modified. In addition also the basic periodicity of one flux quantum is not affected, and it even can be shown that this property is very fundamental holding for a broad variety of conducting systems.
Within the 1D description it is evident that the amplitude of the conductance oscillation (or fluctuation) will always be $< \sim e^2/h$. In the multi-channel case the amplitudes will be affected, in particular via elastic effects. It is of considerable interest that also in this case the amplitude of the conductance fluctuation *again is* $\sim e^2/h$. This behaviour is found to be universal for the quantum diffusive regime, and we discussed it already very briefly at the end of chapter 2. Effectively it seems as if the transmission of the individual channels contributing to the total transmission do not behave independently, but are somehow coupled by a so-called sum rule. This is a very fundamental property of *random quantum systems*.
The next aspect that needs consideration is how the amplitude of the oscillations is affected by external parameters; in particular we will discuss the role of different energy scales in the system, namely the temperature and the applied (measuring) voltage. This will allow us to become acquainted with the role of the Thouless energy in mesoscopic systems (see chapter 1, section 2b).

To see this, we have to evaluate how the interference of an electron depends on its energy. The kinetic energy of the electrons affects their phases via the velocity or the ν-vector acting in the first term of eq. (5.2). From the eqs. (1.17) and (1.18) in chapter 1 it is seen that the *typical* change of energy ΔE that yields a change in phase of π, -which by definition is the Thouless energy E_{Th}-, is given by $\Delta E=E_{Th}=(h/2\pi)/\tau_L$, with τ_L being the time it takes for an electron to travel the (characteristic) size of the system L. So let us compare the interference of electrons at two energies, E and $E'=E+E_{Th}$. If we assume that the electron at energy E leads to *constructive interference* at the output (i.e. $\Delta\phi_1-\Delta\phi_2=0$ in the last term of eq. (5.5)), then for the electron at the energy E' the additional phase shift of π implies that the term $\Delta\phi_1-\Delta\phi_2=\pi$, or this electron wave experiences *destructive interference*. Thus, we find that the transmission probability T through the ring is energy dependent, including its zero-flux phase. This has a drastic effect on the conductance of the ring. The total conductance of the ring follows from the integration over all contributing electrons (and so over the entire relevant energy range ΔE), divided by that range, or $\frac{1}{\Delta E}\int_{\Delta E} T(E)dE$. If the energy range $\Delta E<<E_{Th}$, the transmission $T(E)$ does not depend on energy and consequently this will hold also for the conductance. If however the system is "fed" by electrons that cover a range $\Delta E>E_{Th}$, we find that part of the electrons interferes constructively while an other part does so destructively, thus reducing the overall (flux dependent oscillatory part of the) conductance. Effectively we may divide the relevant energy interval ΔE into $N_b=\Delta E/E_{Th}$ energy "slices" which have a different amplitude with alternating zero-flux phases. This leads to partial cancellation. So in conclusion, *once the energy window of the electrons entering the ring system surpasses the Thouless energy, the amplitude of the oscillatory part of the resistance becomes reduced.* A quantitative analyses shows that the reduction follows $\sqrt{E_{Th}/\Delta E}$. The energy interval ΔE can be controlled by the bias voltage applied across the entrance-exit leads, i.e. $\Delta E=eV_{bias}$, or by the temperature of the electron reservoirs of the leads ($\Delta E\sim 3kT$), and so we envision a bias voltage and a temperature dependence of the AB oscillation amplitudes.

5.2.b. *Experimental results of AB and AAS oscillations.*

Now we are ready to discuss a few experiments on the AB and AAS effect. Figure 5.3 shows a typical sample comprising a small, narrow Au ring fabricated by standard electron-beam lithography and metal evaporation (R.A. Webb et.al, Phys. Rev. Lett. 54, 2696 (1985)). The typical dimensions are 820 nm diameter, 50 nm wirewidth and 40 nm thickness of the Au film. Figure 5.4 shows the results obtained from the measurement of the conductance in dependence of an externally applied magnetic field, at a temperature of 40 mK. The clear, oscillating pattern nicely demonstrates the AB effect in the solid state in the diffusive transport regime to be real. The Fourier transform of the conductance oscillations given in Fig. 5.4b shows a clear peak in "frequency" $1/B$. Taking the surface area enclosed by the loop, this fully complies with the h/e AB oscillation period, i.e. a change of one flux quantum through the loop per oscillation period. Also a

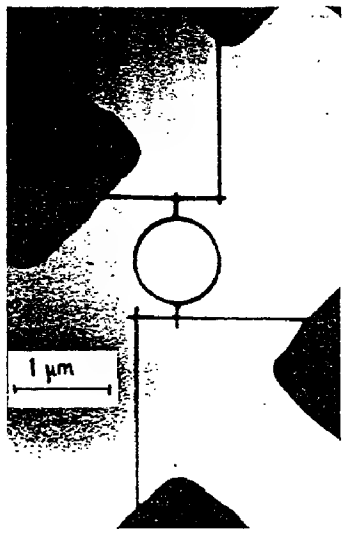
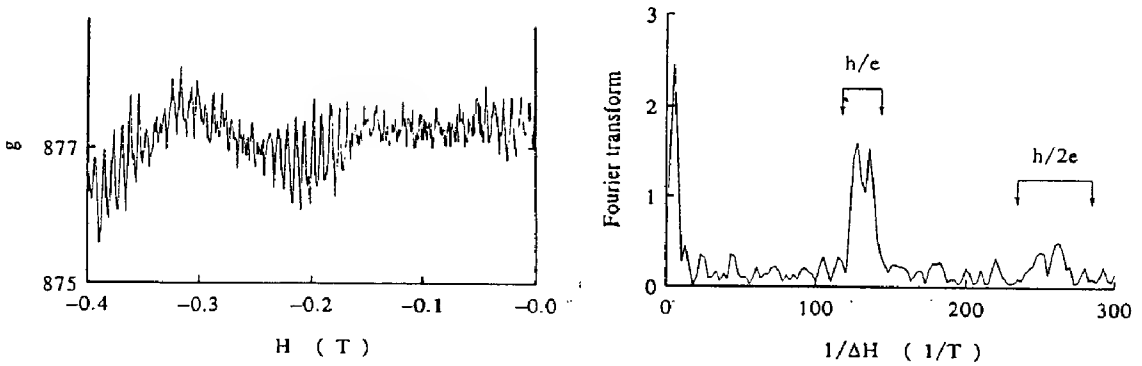


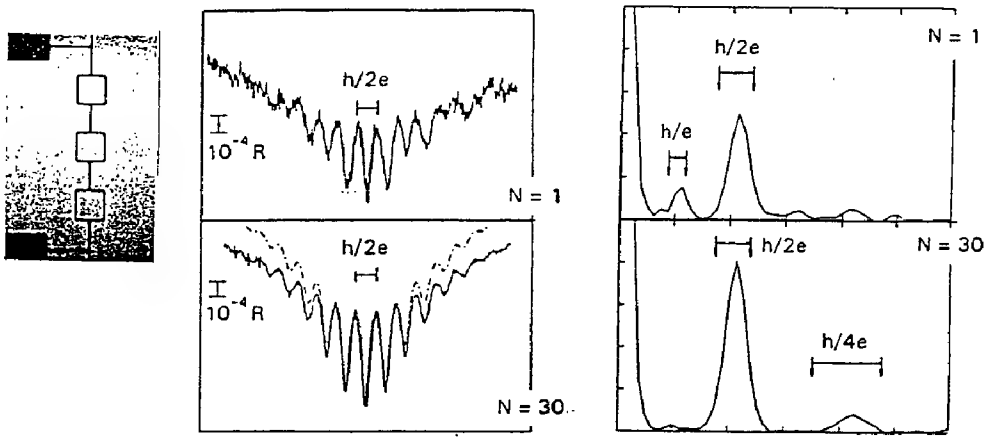
Figure 5.3. A gold ring with four leads (two at the top and left, two at the bottom and right) to study Aharonov-Bohm effects in the quantum diffusive regime.

very weak *h/2e* feature is seen. From more detailed experiments (at large magnetic fields) it became clear that this is a harmonic of the AB oscillation and not an AAS contribution, i.e. not due to time-reversed trajectories. From the discussion on the AB and AAS



a. *b.*

Figure 5.4. Conductance oscillations in the ring of fig. 5.3, clearly demonstrating the AB effect. a: A sample of the data; b: The Fourier transform of the data of a.



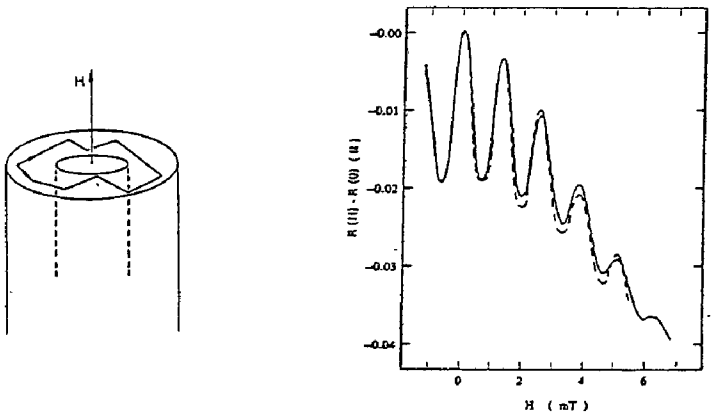
a. *b.* *c.*

Fig 5.5: Ensemble averaging in multi-ring devices. a.: example of a 3-loop structure; b.: resistance data for a single ring (N=1) and the device of N=30 rings; c. Fourier transforms of the data shown in b.

oscillations it became obvious that the first have a phase with a *random value* 0 *or* π of the oscillation at *zero flux*. In contrast, the AAS oscillations, resulting from time reversed path, always will be at a *maximum* resistance at zero flux. This has an immediate consequence on the transport through an ensemble of interconnected rings. For the AB oscillations the total resistance will be a sum of the randomly phased AB contributions and so this will lead to a significant suppression of these *h/e* oscillations. This does not hold for the AAS oscillations. Here the zero flux phase is well determined and so the contributions of all individual rings are added "in phase". This is exactly what is found experimentally.

Figure 5.5 provides data obtained from a series of N=30 loops connected in a (1D) array, compared to a single loop (N=1). Just as an illustration figure 5.5a shows a three-loop device; the 30-loop structure is made in a comparable way. The measured resistance versus the applied field clearly shows an oscillatory shape (Fig. 5.5b). Note that in this case the AAS contribution is larger than the AB part (fig. 5.5b & c), the reason of which we will not discuss here. However, the *h/e* contribution is visible in the single ring while it is fully suppressed in the N=30 ensemble, in accordance with our preceding discussion.

Before closing this section we want to discuss one more experiment, stressing the role of the phase coherence length in these type of experiments. Figure 5.6a shows a peculiar "AB"-configuration, i.e. a long, hollow conducting tube, with the magnetic field directed



a. b.

Figure 5.6. AAS oscillations in a metal-coated thin glass fibre. a.: The sample configuration; b: Resistance oscillations in dependence of the applied field along the tube.

along the fibre. The tube is made by evaporating a thin metallic layer onto the surface of a glass fibre of typically 1 mm diameter. The length L of ~1mm largely exceeds the phase coherence length l_ϕ. This effectively means that we can subdivide the tube *along its length in ~L/l_ϕ phase independent rings*, each "slice" forming one member of the ensemble of ~L/l_ϕ (>>1) individuals. Because of this large number we anticipate that strong AAS oscillations should be present. This is confirmed in the experiment (figure 5.6b), which shows the results of a measurement of the conductance along the fibre.

5.3 Persistent currents in non superconducting systems

In this section we discuss the phenomenon that a steady state current can exist in a ring shaped conductor. From classical electrodynamics it is known that a current can be set up if a (steady) voltage source or a (time dependent) magnetic field is to generate a current in a closed conductor. Instead, in a mesoscopic structure of ring shaped topology a DC current can arise with a constant, time independent magnetic field. Thus, *the current is a thermodynamic equilibrium property* of the system. It arises if the phase of the electron travelling around the ring is preserved after a full revolution, which gives rise to well defined eigenstates. These states, which are the quantumstates of the electronic system of the whole ring, result in a finite circulating current. The value and direction (or sign) of this current depend on the shape of the energy spectrum and on the occupation of the states of the spectrum. In this way it will be affected by the number of electrons residing in the ring and the magnetic flux enclosed by the ring.

Persistent currents are considerably more familiar then it may seem at first sight. Electrons orbiting the nucleus/i of an atom or a molecule in steady state are well known and understood. These stable electron orbits or states are purely quantum mechanical in origin. Effectively these orbiting electrons form persistent currents and lead to a finite magnetic susceptibility, which can be either dia- or paramagnetic, partially depending on the number of orbiting electrons. Some of these aspects will be discussed in more detail in Chapter 11.
A second example of persistent currents was introduced implicitly in chapter 4. In a 2DEG subject to a large magnetic field the electrons will enter the Integer Quantum Hall state. Under this condition edge states will be formed at the crossing of the Landau levels

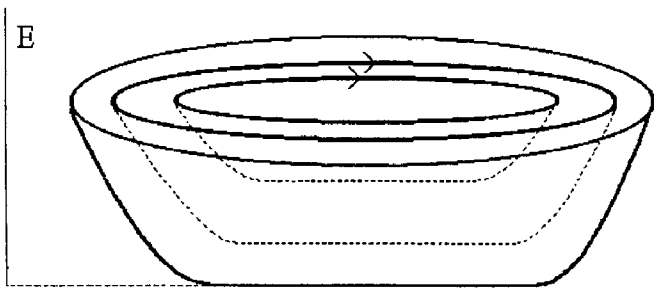


Figure 5.7. Persistent currents in the quantum Hall regime, associated with electrons travelling in edge states along the perimeter of the sample.

and the Fermi energy. Electrons residing in these edge states will travel along the edge of the 2DEG in a direction determined by the direction of the magnetic field. This results in a net transport of electrons along the circumference of the 2DEG, i.e. a persistent current. Fig. 5.7 shows a (circular) section of 2DEG in an energy-position landscape. The (two) Landau levels result in just as many edge states at the Fermi energy and each edge state will contribute to the circulating current.
At a magnetic field B electrons travelling along the edge through a local electrostatic field E have a velocity of approximately $v \cong E/B$, and so *each electron* in an edge state contributes a current

$$i = ev/L \cong eE/LB \qquad (5.8)$$

with L being the length of the circumference of the 2DEG section. The current due to the circulating electron generates a magnetic moment $M=iS$, where S is the area of the 2DEG involved.

Problem: Estimate the magnetic moment for the case of a single edge state, taking $S=1$ μm^2 and a typical value of $E \sim 10^4$ V/m for the electric field at the edge. (Note: to estimate the number of electrons involved assume a linear gradient of the electrostatic potential at the edge). Compare the calculated M with the external field required to establish the IQH state with one edge state.

After these two examples based on preceding discussions in the next sections we will introduce persistent currents in normal conducting rings. In the first subsection, 5.3.a, we start with the basic theory underlying the phenomenon, starting from a ballistic 1-dimensional ring to see the role of quantum eigenstates. This is generalised by introducing scatterers (and thus going from ballistic to the diffusive case) and allowing multiple mode occupation. Also the mechanism of period halving in multi-ring structures will be introduced. Next, in subsection 5.3.b. the presently available experimental results are discussed. We finalise the chapter by briefly discussing the relation of the persistent current problem and the AB effect.

5.3.a. *Theory of persistent currents in rings of normal conductors*

In this subsection we introduce the main aspects of the theory of persistent currents. It will be shown that persistent currents arise because of the existence of well defined quantum eigenstates in the ring. An electron occupying such an eigenstate has a non-zero velocity, which results in a stationary circulating current. Most notably, this current is an *equilibrium property* of the electron in the ring. The magnitude of the persistent current depends on the E-k and/or E-B relation, with B and applied magnetic field that pierces through the opening of the ring. We will see how elastic scatterers in the ring affect the E-k spectrum, and in this way the size of the circulating current.

Figure 5.8 shows a metallic ring with a circumference of length $L=2\pi R$, placed in a magnetic field B such that a flux Φ is enclosed by the ring. The width of the ring is taken to be such as to accommodate just one single mode or subband, i.e. roughly speaking it is approximately half a Fermi wavelength. In this way the system is 1-dimension (1D). We

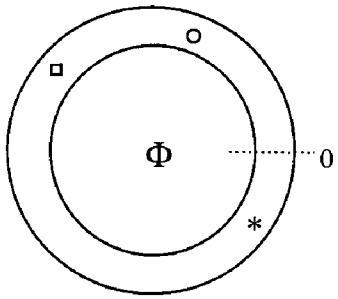


Figure 5.8. Narrow isolated metallic ring with elastic scatters.

assumed that electrons travelling around the ring preserve their phase during at least one round-trip, i.e. $L<l_\phi$. The ring may contain a few elastic scatterers, which we initially assume to be weak. Electrons travelling around the ring will experience the same conditions again (e.g. the potentials due to the local scatterers) each time they have completed one round-trip. Stated differently, they will behave as if they are travelling in a *periodic potential*, with a period of length L. Fig 5.9 shows this periodicity more explicitly by "unfolding" the ring into a strip with segments of length L each.

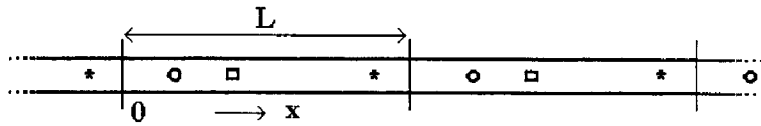


Figure 5.9. The periodic system resulting from cutting, unfolding and repeating the ring structure of Fig. 5.8.

The system we thus obtain is completely identical to the periodic structure experienced by an electron in a solid state crystal lattice. This allows us to use the same techniques to evaluate its properties. From quantum mechanics we know that this periodic motion leads to the formation of eigenstates, with associated eigenvalues for the energy, wavevector and velocity. To simplify the discussion at this stage, we assume that the scatterers are absent; we will re-introduce them later again. At zero magnetic field we immediately can deduce the eigenvalues for the k-vectors:

$$k_p = \pm p\frac{2\pi}{L} \qquad (p=0, 1, 2,...) \qquad (5.9)$$

The quantum number p labels the states. As discussed before in section 5.2, applying a magnetic field will affect the phase of the electron via the vectorpotential (eq. 5.1), resulting in an additional phase shift $\Delta\phi$ depending on the flux Φ enclosed by the ring (eqs. 5.2 and 5.3)

$$\Delta\phi = 2\pi\,\Phi/\Phi_0 \qquad (5.10)$$

with Φ_0 denoting the flux quantum for single electrons, $\Phi_0 = h/e$. Evidently this phase shift will affect the eigenvectors k_p, resulting in a modified condition for establishing a bound state

$$k_p(\Phi)L - \Delta\phi = p2\pi \tag{5.11a}$$

or

$$k_p(\Phi) = (p + \frac{\Phi}{\Phi_0})\frac{2\pi}{L} \tag{5.11b}$$

From this condition for the quantisation of the momentum we immediately can derive the velocity and the (kinetic) energy of the states of quantum number p

$$v_p(\Phi) = \frac{\hbar}{m}k_p(\Phi) = \frac{\hbar}{m}(p + \frac{\Phi}{\Phi_0}) \tag{5.12}$$

$$E_p(\Phi) = \frac{\hbar^2}{2m}k_p^2 = \frac{\hbar^2}{2m}\frac{4\pi^2}{L^2}(p + \frac{\Phi}{\Phi_0})^2 \tag{5.13}$$

Fig. 5.10 shows the resulting E-Φ-diagram, with the quantum number p labelling the successive parabolas. Note the similarity to the E-k diagram as obtained in the common solid state case. As anticipated the figure shows periodicity in the number of flux quanta contained within the area of the ring.
The current carried by the circulating electron can be calculated simply by

$$i_p = \frac{dQ}{dt} = \frac{-e}{L/v_p} = -\frac{ev_p}{L} \tag{5.14}$$

The current carried by such a bound state can also be obtained in a more general way. To

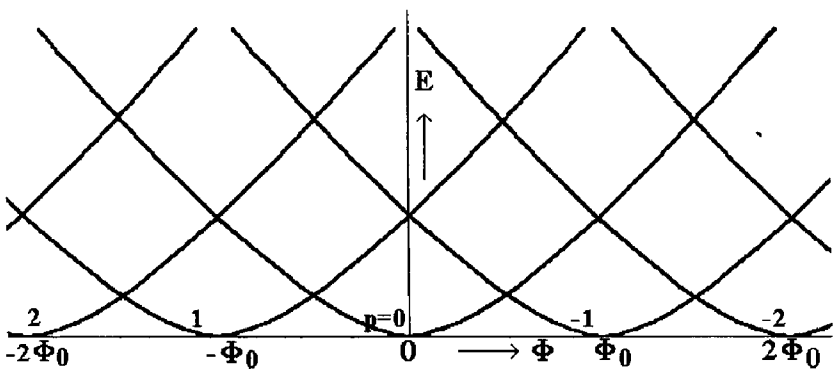


Figure 5.10. The E-Φ -diagram for electron states in the 1D ring of Fig. 5.8

see this we calculate the derivative of the kinetic energy to the flux Φ

$$\frac{dE_p}{d\Phi} = \frac{\hbar^2}{m} k_p \frac{2\pi}{L\Phi_0} = \frac{h}{L\Phi_0} v_p = \frac{e}{L} v_p(\Phi) \qquad (5.15)$$

Noting that the two right-hand side expressions of eqs. (5.14) and (5.15) are equal, we conclude that the current can be written as

$$i_p = -\frac{dE_p}{d\Phi} \qquad (5.16)$$

This equation is the well known thermodynamic relation between the magnetic moment M ($=iS$) and the derivative of the (Gibbs free) energy to the magnetic field B ($= \Phi/S$), with S denoting the area enclosed by the circulating current i which generates the magnetic moment M. Note that i_p is the current contributed by a *single electron in a specific state labelled by p*. As commonly the magnetic field employed in these experiments is small, the spin degeneracy will not be lifted and so eq. (5.16) has to be multiplied by the factor g_s=2 to obtain the total current per fully occupied level (i.e., two electrons, one spin up and one spin down).

Before continuing to show how one obtains the total current resulting from the contributions of all the electrons in the ring we first want to discuss the effect of the elastic scatterers as shown in Fig. 5.8. The effect of each scatterer that it introduces a non-zero localised electrostatic potential in the system. Given the ring shape, this leads to a *periodic potential* for the electrons as shown in the unfolded system of fig. 5.9. From introductory solid state physics we know that the non-zero value of the periodic potential results in Bragg reflection of the electron waves and *the formation of energy gaps and bands*. This is shown in fig. 5.11. Because of the formation of energy bands a new set of quantum numbers n is introduced, which enumerates the bands starting from the one that

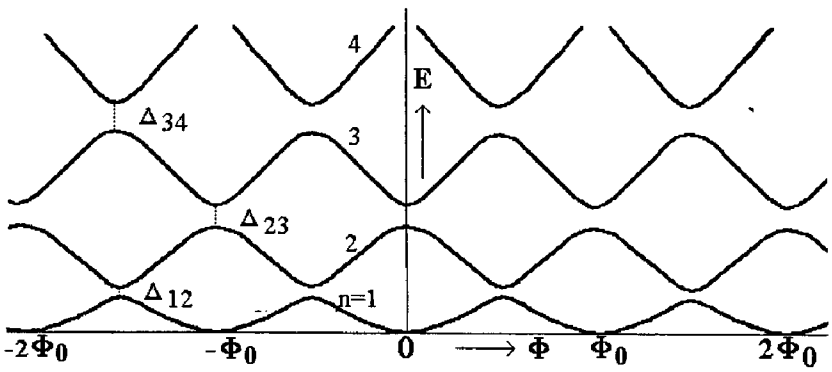


Figure 5.11. The finite potentials of the elastic scatterers in the ring of Fig. 5.8 lead to the formation of gaps in the parabolic E-Φ -diagrams of Fig. 5.10.

is lowest in energy. The size of the gaps can be derived in a simple way from the potential $V(x)$ along the ring (x represents the position coordinate running along the ring). As $V(x)=V(x+L)$ is periodic, it can be rewritten as a Fouriersum by (see e.g. Kittel, Introduction to Solid State Physics)

$$V(x) = \sum_{k_p} U_k \exp(ik_p x) \tag{5.17}$$

This allows one to calculate the size of the gaps as

$$\Delta_{12} = 2\,U_1,\ \Delta_{23} = 2\,U_2, \ldots. \tag{5.18}$$

So, strong potential variations $V(x)$ generally will lead to an increase in the gaps with a simultaneous reduction of the width of the bands. Stated differently, an increase in disorder (or: a reduction of the elastic mean free path) leads to larger gaps and narrower bands. Equivalently, crossing from the nearly free electron case with its weak scattering to the tight binding regime with strong scattering leads to an increase in the localisation of the electrons, and we anticipate that this will be accompanied by a reduction of the average velocity of the electrons in the states, i.e. the current. We return to this below.

Now all the ingredients to calculate the *total* current in the ring are available, by appropriately summing all individual currents in each band. For the case of the parabola of the clean (impurity free) case, each state contains two electrons at maximum, due to spin. Equivalently, for the case of elastic scattering, each band only can accommodate two electrons. So, we assume 2N electrons in the ring occupying N bands (n=1,2,3,..,N). Referring to fig. 5.11 it is obvious that the maximum current (or, the maximum slope of E versus Φ, see eq. (5.16)) occurs at $\Phi \cong \Phi_0/4$. For the case of small gaps, i.e. for weak disorder, the *typical maximum* current in band n for this value of Φ can be obtained from eqs. (5.11b, (5.13) and (5.16) as

$$i_n \approx (-1)^n g_s \frac{eh}{m}\frac{1}{L^2}((n-1)+1/4) \tag{5.19}$$

The prefactor $(-1)^n$, indicating the sign of the derivative of E with respect to Φ, alternates for successive bands with quantum numbers n, a property which is immediately evident from figure 5.11. Consequently, *currents of successive bands largely cancel, yielding a sum that is approximately equal to the contribution of a single band*, say the topmost band with label N. Thus as an order of magnitude estimate for the total current of N filled bands at $\Phi \cong \Phi_0/4$ we find

$$|I_N| = |\sum_{n=1}^{N} i_n| \sim |i_N| \approx g_s \frac{eh}{m}\frac{1}{L^2} N \approx 2e\frac{v_N}{L} \tag{5.20}$$

Problem: derive the last "equality" of eq. (5.20)

The argument for the *current cancellation of successive bands* is independent of the scattering strength. So, the "equality" $|I_N| \sim |i_N|$ provides a way to evaluate the *total current* for any scattering strength, assuming the bandstructure is known (which is by no

means obvious!) so that the current in the N-th band, $|i_N|$, can be evaluated using eq. (5.16).
From eq. (5.16) one also can understand how the current will be affected by an increase in the disorder in the ring. Stronger scattering will lead to larger bandgaps (eq. (5.18)) and consequently to narrower bands. These narrower bands have a smaller slope of energy with respect to the flux, resulting in a reduction of the current per band (eq. (5.16)). From the first part of eq. (5.20) it is then obvious that the total current will be reduced compared to the nearly-free electron case of small scattering.

Problem: Calculate the total current for a metallic ring of L=1 μm in the weak scattering regime. How many electrons and bands are involved? (Answ: ~ 1 μA)

To obtain a rough estimate of the persistent current in the case of increased scattering we do not need to evaluate the bandstructure of a disordered ring. Effectively an increase of the scattering makes the electron transport diffusive by the time the elastic scattering length l_e becomes considerably smaller than the length of the circumference of the ring, L. For a strip of size L carrying N_{ch} modes we found for the conductivity within the Drude model (see eq. (2.5))

$$\sigma_{Drude} = g_s \frac{e^2}{h} N_{ch} \frac{l_e}{L} \tag{5.21}$$

i.e., it is reduced by a factor l_e/L as compared to the ballistic case. Note that the reduction of the conduction has an elastic, non-dissipative origin and will not affect the phase *coherence* in the system! The same factor l_e/L can be introduced in eq. (5.19) to obtain an *order of magnitude* estimate for the persistent current in the diffusive regime, yielding

$$i_n \approx (-1)^n \frac{l_e}{L} \frac{eh}{m} \frac{1}{L^2} ((n-1)+1/4) \tag{5.22}$$

A word of caution is appropriate here. Including the effect of elastic scattering by the introduction of the prefactor l_e/L should be taken as a crude first order guess. In particular for the case of few channels (so including the 1D case we have been discussing here) its validity is rather limited!

Now that we have discussed the effect of elastic scattering in some detail, the question arises on the effect of inelastic processes on the magnitude of the persistent currents. This problem can be approached largely along the same lines as in the preceding section 5.2 on the AB effect. Inelastic scattering will reduce the phase coherence length l_ϕ and time τ_ϕ, and consequently will decrease the lifetime of the eigenstates of the ring. As usual this leads to a broadening of the energy levels of these states. By the time the broadening of the levels becomes comparable to the typical distance between the bands the bandstructure will become disrupted and the current will become suppressed significantly. This suppression will be given by the same factor as before (eq. (5.6a))

$$i_n = i_{n0} \exp(-\frac{L}{l_\phi}) \tag{5.23}$$

Before continuing the discussion of the theory of persistent currents we first take a small sidestep to some experimental details. This is in order to estimate the magnitude of the total current in a ring, and how it has to be measured.
Experimentally, in the case of a ring fabricated by metal evaporation, the typical thickness of the metallic sheet will be some 30-100 nm (see also Appendix A), containing typically 10^6-10^8 electrons. Assuming diffusive scattering at the substrate-metal interface, l_e will be of the same order of magnitude. Consequently in a ring of $L\sim$ 1 μm the persistent current will be reduced by a factor $l_e/L\sim$10-30 compared to the ballistic case (see eq. (5.22). As we have found the typical current for the ballistic case to be <~1 μA, we end up with $|I_N|<\sim$0.1 μA in the diffusive regime: a rather small current! This current not only is rather small, but in addition it can *not* be measured in the conventional way by "cutting" the ring and including a current meter in the circuit. The inclusion of such meter (based on classical, dissipative phase breaking mechanisms!) will disrupt the quantum eigenstates of the ring and thus suppress any persistent currents associated with them. The only method to determine these currents is by measuring the magnetic moment $M=iS$ resulting from the circulating currents, e.g. by employing a magnetometer like a SQUID. We will return to it later in the chapter when we discuss the experimental results (see also Appendix B). The matter of experimental accessibility is brought up here to pose the obvious question: is there a way to increase the total magnetic moment resulting from persistent currents? Two possibilities may come to one's mind: employing a ring accommodating more than one channel; or using a number of separate identical rings.

Let us first consider the problem of a single ring containing more than one channel, say M channels (or modes/subbands) enumerated by m=1, 2, 3, ... In the case of a complete absence of elastic scattering the M channels will behave independently, with each channel forming a set of bands. Assume mode m to contain N_m filled bands, i.e. $E_{N_m} = E_F$ and $E_{N_m+1} > E_F$. Due to the complete independence of the modes a multichannel equivalent of the bandscheme of Fig. 5.10 is results, and each subband will show up as a set of parabolas, of course again periodic in Φ_0. However, each set will have its bottom at an (increasingly larger) subband minimum E_{m0}, which is as usual determined by the associated quantised perpendicular wavevector k_m (see chapter 2). For each subband m, containing N_m filled bands, the current per state can be calculated from eq. (5.16), while the total net current *per subband* is obtained from eq. (5.20). To calculate the total current we have to sum the contributions of all individual subbands. In doing so we again have to consider the argument of partial cancellation, this time resulting from the prefactor $(-1)^{N_m}$, which will be (rather randomly) positive or negative for each individual channel with N_m filled bands. As a consequence the resulting sum again will be similar to that given by eq. (5.20), i.c. of the order ev_{max}/L, with v_{max} denoting the velocity of the highest lying occupied state of the lowest subband, or stated differently, from the state with the largest kinetic energy or velocity.
So, in conclusion, increasing the width of the ring to accommodate more than one channel, does *not* increase the net typical persistent current in the ring, despite the increase of the number of electrons in the system!

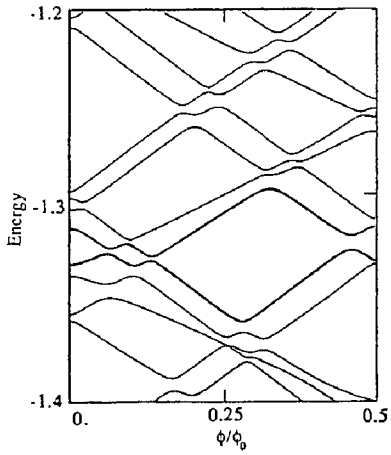


Figure 5.12. Part of the energy-flux bandstructure for a disordered ring in the weak scattering limit; it contains M=10 lateral subbands.

Using approximative methods it is possible to calculate the bandstructure of disordered systems of the nature we have been discussing. Although we will not go into any detail of such calculation, an example of the complexity of such bandstructure of a multichannel ring is shown in figure 5.12 (H. Bouchiat et.al, J. Phys. France 50, 2695 (1989)). Here the limit of weak scattering is taken. The ring contains M=10 subbands.
Only the flux range 0-$\Phi_0/2$ is shown, as the current will be symmetric around Φ_0=0 and periodic in one flux quantum Φ_0. In the diagram the opening of (small) gaps due to the weak disorder is clearly visible. Note that none of the states ever *crosses* another state: the gaps effectively make all states to avoid each other! This very general behaviour of eigenstates in complex systems is called *level repulsion*, something we will come back to in Chapter 9. Note also that no longer the maximum slope (and so the maximum current) occurs at $\Phi_0/4$: this is a direct consequence of the mixing of the individual subbands by the elastic scattering.

Now we want to address the second possibility raised with respect to increasing the total magnetic moment due to persistent currents, i.e. by combining a number of (nominally) identical rings. Here we will again encounter the effect of *period halving*, i.e. the magnetic moment due to the total circulating current shows a periodicity $\Phi_0/2$, similar to what was found in the AB/AAS effect in a multi-ring configuration in the preceding section 5.2 (see fig. 5.5).
Metallic rings of micrometer diameter typically contain at least 10^6 electrons. If an ensemble of nominally identical rings is fabricated the number of electrons residing in one particular ring, however, may differ considerably. Depending on the quality of the fabrication process, variations up to ~ 1% are realistic, which is equivalent to at least 10^4 electrons. In addition the details of the bandstructure of each ring will differ as the actual positions of the scatterers in each individual ring will be different. Noting that the difference of *one single electron* may already *reverse* the direction of the current (see eq. (5.19)), it is clear that these much larger variations will have a profound effect. So, nominally identical rings will differ widely when it comes to thermodynamic equilibrium

electronic properties. Let us look to the consequences for the magnetisation of an ensemble of rings in a little more detail.
In order to keep the discussion relatively simple we assume an ensemble of three 1D rings in the weak scattering regime, each of size L. Each of the R=3 rings, indicated by r=1, 2 and 3, contains a different number of electrons N_r and again for simplicity we assume the

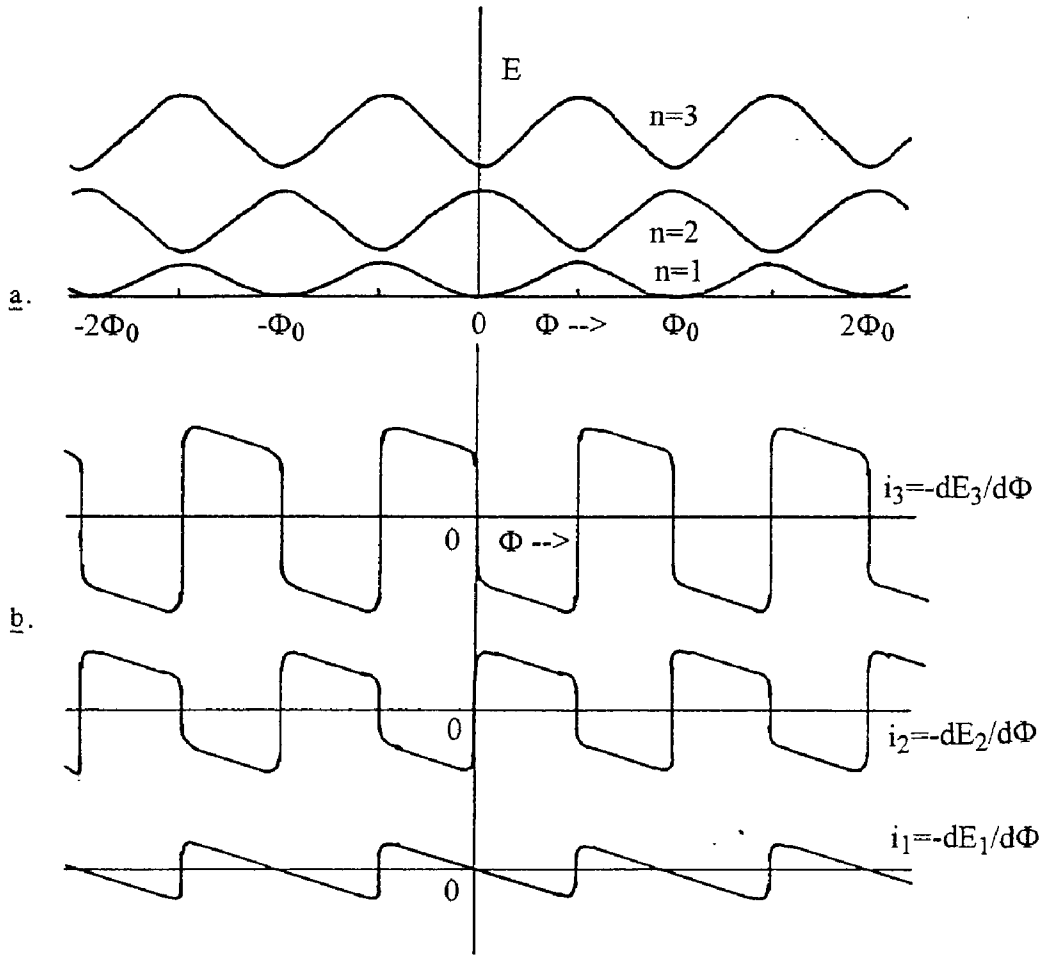


Figure 5.13. Persistent currents in an ensemble of 3 rings, containing 1, 2 and 3 electrons respectively; a. the energy diagram showing 3 bands: for the ring containing 3 electrons all bands are occupied, bands 1 and 2 will be filled for the ring with 2 electrons, while only the lowest band is filled for the ring containing a single electron; b. the derivative of energy versus flux for each of the bands, providing the persistent current per (filled) band.

first ring to contain N_1=1 electron, the second N_2=2 and the third N_3=3; we do not take the spin into account, so g_s=1. Fig. 5.13a shows the resulting bandstructure E-Φ, drawn for ring r=3, with the three bands involved. Note that the *details* of the bandstructure of the rings 1 and 2 will be different (e.g. the gaps may differ in size), but *globally* they will behave similarly. For ring 1 (containing just one electron) only the lowest band is filled, while for the second ring bands 1 and 2 are occupied. In Fig 5.13b the currents $i_{3,n}(=-dE_{3,n}/d\Phi)$ for ring 3 are shown, for each individual band n=1, 2 and 3. Note again that the currents for ring 1 and 2 may differ in magnitude because of the difference in e.g. the size of the gaps, but the *global* shape of $i_{1,n}$ (n=1) and $i_{2,n}$ (n=1,2) will be rather similar.

The next step is to calculate the total current per ring as $I_r = \sum_n i_{r,n}$ for the 3 rings r=1,2 and 3 (=R). This is shown in figure 5.14. This figure also shows the mean value $\langle I_R\rangle=(I_1+I_2+I_3)/R$ of the 3 rings. From the figure it is clearly seen that the mean current shows a pronounced *halving of the period, i.e. a* $\Phi_0/2$ periodicity. This is a characteristic feature resulting from the averaging process; it becomes increasingly pronounced if

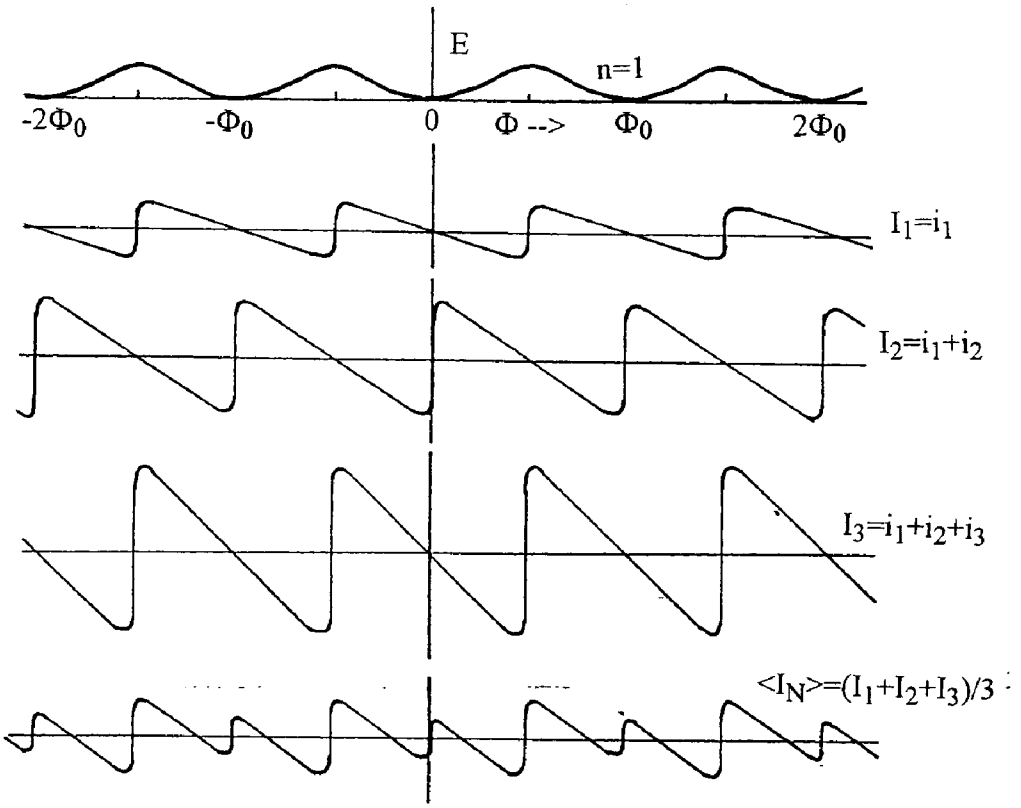


Figure 5.14. The calculated total persistent current for each ring, I_1, I_2 and I_3, and the resulting mean value of the currents in the three rings, $\langle I_3\rangle$.

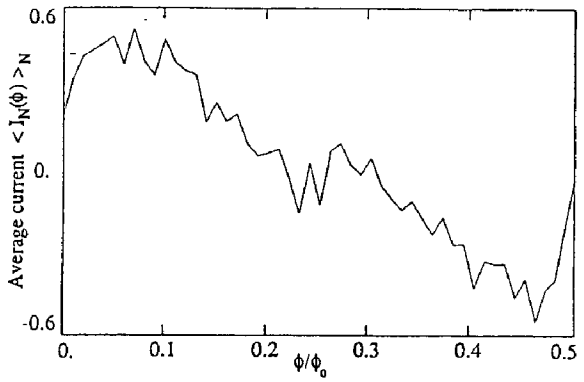


Figure 5.15. Mean persistent current for a large ensemble of similar rings. The number of filled bands for each ring, N_r, is distributed evenly between 150 and 250.

the averaging is performed over larger ensembles of rings (or larger R). Note also that the magnitude of the mean current is comparable to that of the current of a single ring.
Now we are ready to evaluate the total magnetic moment due to the circulating persistent currents in all three rings. If S is the area per ring, this moment is given by
$M= M_1+M_2+M_3= S(I_1+I_2+I_3)= S*R*<I_R>$.
This result leads to two important conclusions. First, *the total magnetic moment scales linearly with the number of rings*, which means that it is experimentally useful to employ a set of rings. Second, the total signal for the set of nominally equal rings *shows period halving*, contrary to the case of a single ring where the persistent current shows oscillations periodic in a full flux quantum.
Figure 5.15 shows the results of a calculation for a set of rings in the limit of weak disorder; the number of filled bands per individual ring ranges between 150 and 250 [H.Bouchiat et.al.]. Clearly in this much more realistic case the characteristic period halving which we have found in the simple example is evident, stressing the universality of this phenomenon.
The suppression of the *h/e* component in the persistent current (or magnetic moment) while preserving the *h/2e* component in averaging over a number of independent rings, basically has the same origin as the survival of the AAS component for the case of the multi-ring AB/AAS conductance as discussed around figure 5.5. It originates because the phase of the half flux quantum component always has a specific behaviour around zero flux. This symmetry forces the individual contributions to sum "coherently", while the one flux quantum contributions add randomly.

Problem: discuss the case that the sizes of the rings contained in the set differ: consider in particular the effects on the amplitude in dependence of the applied flux.

5.3.b. *Experimental results of persistent currents in single and multiple rings*

In this subsection we will consider three cases: a single metallic ring in the diffusive regime, a single semiconductor ring in the (quasi-)ballistic regime, and an experiment on a set of approximately 10^7 metallic rings. It should be noted that this sequence is historically incorrect, as the experiment on the set of rings preceded the single metal and semiconductor ring cases, basically due to the large experimental difficulties for the single ring case.

1. Single metallic ring in the diffusive limit.

In 1991 Chandrasekhar et.al. (Phys Rev. Lett. 67, 3578 (1991)) presented their experimental results obtained on a single gold ring. The ring had a diameter of 2.4 μm, a width of 90 nm and a thickness of 60 nm (Fig 5.16b). A rough estimate based on the width and thickness indicates that some $M=10^5$ transverse modes are occupied in this device. From resistance measurements on a simultaneously evaporated gold strip the elastic mean free path is found to be ~70 nm and from weak localisation measurements (see chapter 2, section 3) on a simultaneously evaporated strip a phase coherence length l_ϕ~ 10 μm (at T=50 mK) is obtained, the last one depending particularly on the density of magnetic impurities in the gold. Fig 5.16a shows a (strongly simplified!) diagram of the set-up for the measurement.

It consists of a SQUID magnetometer, two coils of superconducting wire to pick up the magnetic moment (one of which contains the ring), and two field coils to generate the magnetic flux through the ring (see also Appendix B). The coils are configured to form a so called (first order) magnetic gradiometer, allowing the discrimination between magnetic signals from the ring relative to those from the environment. Any magnetic signal induced in the two pick-up coils *simultaneously* is suppressed as these are *counter wound*, while the signal due to the magnetic moment of the ring induced in a single coil is transferred to the SQUID detector.

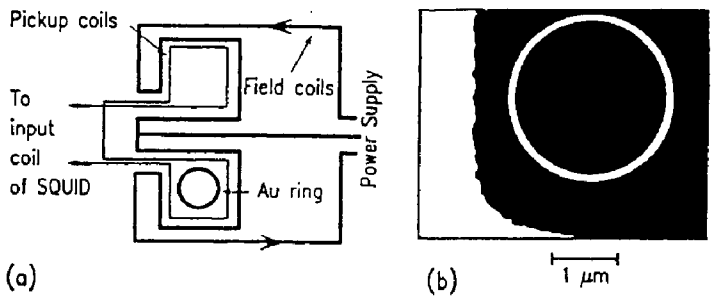


Figure 5.16. a. Experimental configuration to measure the magnetic moment resulting from persistent currents. The field coils generate the flux; the pickup coils sense the magnetic field resulting from the applied flux and the contribution from the magnetic moment of the ring. This total signal is fed to a SQUID magnetic detector. As the pickup coils are counter wound, they form a (first order) gradiometer, leaving the (unbalanced) signal from the ring while suppressing the (balanced) signal from the applied flux. b. An SEM image of the single Au ring of ~ 2.4 μm.

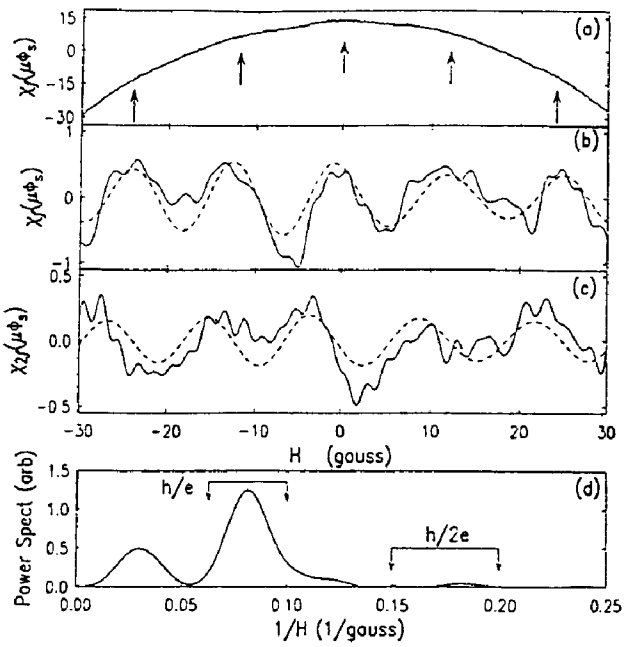


Figure 5.17. The first (a, b) and second order (c) derivatives of the magnetic moment of the Au ring. a shows the raw data directly derived from the SQUID detector. Subtracting a (fitted) second order polynomial results in b and c (solid lines). d shows the Fourier transform of the data of b, with a clear peak at a period of single flux quantum.

To enhance the sensitivity the measurement is performed by adding an AC modulation to the flux applied by the field coils, i.e. $\Phi = \Phi_{DC} + \Phi_{AC} = S*B_{DC} + S*B_{AC}*\sin(2\pi*fmod*t)$, with $S*B_{DC} \sim \Phi_0/10$ typically. The signal induced in the SQUID can be synchronously detected in a lock-in detector, not only at the modulation frequency *fmod*, but also at its harmonics $n*fmod$ (n=2,3,4,..). Detecting at the n-th harmonic effectively means a measurement of the n-th order derivative of the total SQUID signal with respect to the applied flux $S*B_{AC}*\sin(2\pi*fmod*t)$. For more details one should consult Appendix B.

Fig 5.17 shows the results obtained by detecting at *fmod* and *2fmod*, in dependence of the applied DC magnetic field B_{DC} given in Gauss (B=1 T=10^4 Gauss). Fig 5.17a shows the total signal obtained at the modulation frequency. It shows a parabolic behaviour superimposed with weak oscillations. Subtracting a (least square fit) second order polynomial results in Fig 5.17b, displaying the oscillation more clearly. Note that, despite extensive signal averaging (a number of independent measurements were combined to obtain this trace) to enhance the oscillatory signal, the ratio of signal-to-noise is only ~5:1! This stresses the difficulties encountered in this type of experiments. Fig 5.17c shows the signal obtained at the second harmonic *2fmod*. To demonstrate the occurrence of the oscillatory component in the signal shown in Fig. 5.17b its Fourier spectrum is displayed in Fig. 5.17d. A clear peak at a "frequency" $1/\Delta B$=800 T^{-1} is evident. Taking into account the area of the ring this corresponds to a single flux quantum *h/e* being added to the ring per oscillation. This is in full agreement with the theory discussed before.

Problem: check the statement on the period using the given diameter of the ring.

In contrast to the good agreement of the period a serious problem arises if the amplitude is considered. Without going in all the details of the experiment one finds that the amplitude of the measured signal is approximately 30-50 times *too large* as compared to the estimate based on eq. (5.22). We will return to this point briefly at the end of this section.

2. Single semiconductor ring in the quasi-ballistic regime

The ring is realised in a 2DEG in a GaAs/AlGaAs heterostructure, obtained by dry ion beam etching (D. Mailly et.al., Phys. Rev. Lett. 70, 2020 (1993)). The bulk 2DEG has a mobility of ~100m^2/Vs and an electron density of ~3.5*10^{15} m^{-2}, yielding a Fermi velocity v_F~2.5*10^5 m/s and an elastic mean free path l_e~ 10 μm (see chapter 1). With an effective width of ~160 nm (taking into account the depletion resulting from the etched

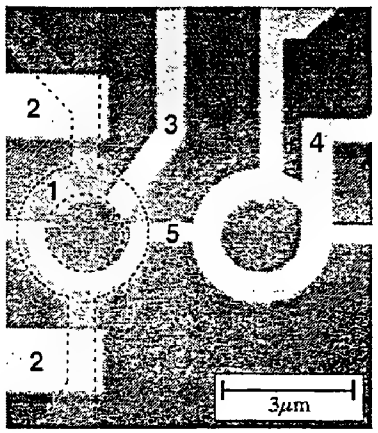


Figure 5.18. Sample layout of the ring used to measure the persistent current in a 2DEG semiconductor. (1) denotes the etched ring with gates (2) (to allow AB measurements) and (3) (to allow suppression of the persistent current. (5) is (a part of) the SQUID

walls) approximately M=8 modes will be occupied in the ring. Apart from some details the experimental approach strongly resembles the preceding one, again employing a SQUID gradiometer to detect the magnetic moment of the ring. Fig 5.18 shows the sample configuration, with only the first part of the SQUID put into place. Note the additional gate structures (2) and (3) in the sample. The first gate pair allows the

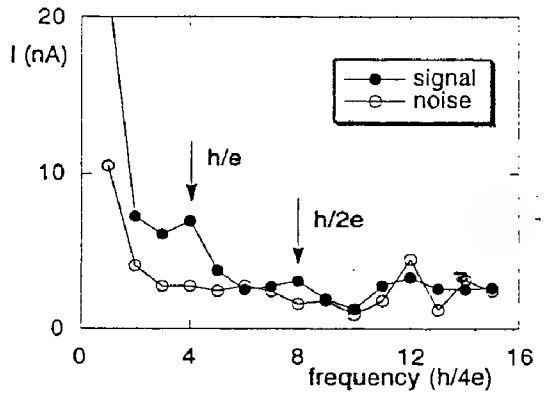


Figure 5.19. Measured persistent current in the 2DEG ring of Fig. 5.18. The figure shows the Fourier tranform of the magnetic moment detected by the SQUID, expressed in terms of an equivalent current.

connection of the ring to external leads to perform AB measurements to be done. Gate (3) controls the current flow in the ring: with a negative voltage applied to it the persistent current can be suppressed, allowing the measurement of the magnetic moment with and without the persistent current flowing. Figure 5.19 shows the results obtained.
Two rather weak peaks can be seen in the Fourier spectrum, with the *h/e* peak being most pronounced. It is of considerable interest to note that in this case, where the electrons in the ring behave approximately ballistically, the measured amplitude of the persistent current of ~5 nA is in *good agreement* with the theoretical prediction based on eq. (5.20) or (5.22). We will discuss this further after presenting our last experiment.

Problem: check this statement on the value of the current.

3. Multiple metallic rings in the diffusive limit.
The third experiment, -which is historically the first of the three we discuss here-, concerns a set 10 million (nominally) equal copper rings and is due to L.P. Levy et.al. (Phys. Rev. Lett. 64, 2074 (1990)). Also in this case the set-up largely resembles the one discussed around figure 5.16a. The result obtained by detecting at the third harmonic $3*f_{mod}$ is presented in Fig. 5.20. Although only a limited number of data points is

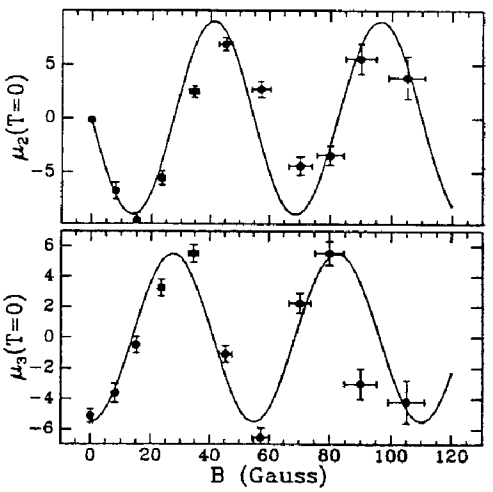


Figure 5.20. Measured magnetic moment of an ensemble of 10^7 copper rings, showing persistent currents. Both the signal at 2f and 3f is shown. As ΔB=130 Gs corresponds to 1 flux quantum through a ring, period halving is evident.

available a clear oscillation is visible. Noting that a single flux quantum through a ring corresponds to 130 Gauss, it is clear that the oscillation shows a period of *half a fluxquantum*, which is nicely in accordance with the prediction given for the multiple-ring case. Concerning the amplitude the evaluation of the magnitude of the magnetisation again leads to a size that is considerably larger than predicted via eq. (5.22), in this case by a factor of approximately 30!

From the three experiments discussed we conclude that good agreement with theory exists on the periodicity of the magnetic moment, with Φ_0 for a single ring and $\Phi_0/2$ for the case of an ensemble of rings. A much less favourable picture arises for the amplitude of the persistent currents. For the ballistic, weak elastic scattering case with only a small number of modes in the ring encountered in the second example of the high-mobility

2DEG, where $l_e/L \sim 1$ or even larger, agreement with theory is good. However for metallic rings, with the electrons moving diffusively, the measured amplitudes are much too large, the discrepancy being at least as large as one order of magnitude! This serious problem is not solved at the moment this text is updated (mid 1997). Various additional mechanisms have been introduced, mostly based on the regulating influence of electron-electron interaction, but none has been completely successful up to now.

5.4. Final remarks

In the preceding two sections 5.2 and 5.3 we have introduced two phenomena that originate from the interference of electron waves in ring-shaped solid state structures. Both these phenomena show a periodic behaviour, with a fundamental period given by the change of the flux that penetrates through the area enclosed by the ring by one flux quantum $\Phi_0=h/e$. In both cases phase coherence is a crucial requirement for the effect to arise, as otherwise interference effects are suppressed.
In discussing the AB and AAS effects in section 5.2 we described the effect as the coherent transmission (and reflection) of electron waves through the ring, allowing such wave to split at one lead-ring junction and reunite at the other. By taking into account the phase shift induced by the magnetic field (via the vector potential) a periodic modulation of the total transmission probability is found, leading to an oscillatory dependence versus the applied flux of the conductance.
In evaluating the persistent currents in section 5.3 we took a different view. Here we assumed the ring to form eigenstates, the energies of which were shown to be flux dependent, again via the vector potential affecting the k-vector and in this way the phase acquired during one roundtrip. The equilibrium state current transported in a state derives from the flux dependence of the energy of that particular state.
Basically, also the AB and AAS effect could have been introduced by starting from the eigenstate picture. In this picture the system is represented by having the states in the leads couple to the eigenstates of the ring. Pictorially this implies that one has to evaluate the transmission coefficient for a certain state in the left lead to a certain state in the right lead via one eigenstate of the ring. After summing over all (occupied) states in the input and output leads one finds the total transmission in going from left to right, i.e. the total conductance. Evidently, while the eigenstates of the ring depend on the flux through its central area, the overall coupling will depend on it, and this will again be periodic in the flux quantum. One may denote this as *resonant "tunnelling"* through the eigenstates of the ring.

General references to the subject:

1. "The Feynman Lectures on Physics", (Addison-Wesley, 1964), Vol 2, pp 12-15
2. The Aharonov-Bohm effect: Physics Today, January 1986, p 17
3. Persistent Currents: Physics Today, April 1992, p 17

6. Transport in normal-superconductor systems

Keywords: superconductor; energy gap; S-N interface; Andreev reflection; phasecoherence; S-N-S systems;

6.1 Introduction

This chapter describes the transport of electrons in systems containing normal conductors as well as superconductors. In particular it concentrates on effects in close proximity to the normal superconductor interface. We will see which lenght- and timescales are of importance and how these affect the electronic properties of the system.

To understand the physical mechanisms leading to the effects introduced here only a limited knowledge of the properties of a superconductor is required. Just three aspects are needed for our discussion: the development of an energygap in the energyspectrum of the normal electrons; the occurrence of Cooperpairs consisting of pairs of electrons; and the formation of a macroscopic quantum state which is occupied by these Cooperpairs, described by a single macroscopic wavefunction.

In a normal conductor, either a (semi-) metal or a (degenerate) semiconductor, the partially filled energybands lead to the existence of a finite density of states (DOS) at and around the Fermi energy. The availability of unoccupied energystates close to the Fermi energy provides the phasespace required for electrons to scatter, in this way allowing electronic transport in the solid state.

These main features of conductors can be completely understood assuming the electrons residing in the conduction band to behave independently. The first interaction which may be considered affecting this independent electron picture is the Coulomb or electron-electron (e-e) interaction. Obviously this interaction is repulsive. Under certain conditions a more complicated interaction may occur additionally. Without stressing the details, this process acts employing phonons as an intermediate. One electron moving through the solid deforms the crystal lattice formed by the ions (quantummechanically speaking: it emits a phonon) via the Coulomb interactions. A second electron moving through the same region of space will experience the deformed lattice, as this results in a local change of the electrostatic potential generated by all nearby ions (in quantum terms: it absorbs a phonon). Under certain conditions this phonon mediated interaction of the two electrons resulting from the lattice deformation is found to be attractive. If this (indirect) attractive interaction is stronger than the (direct) Coulomb repulsion, the net interaction will be attractive. Below a certain critical temperature T_c this attractive interaction may result in the formation of electron pairs, as such a state is energetically more favourable.
These new composite particles, called Cooperpairs (CP), consist of two electrons of opposite k vector and opposite spin, so the effective momentum k=0.

Cooperpairs assemble in a peculair way. They all form one single quantum state, characterised by a single wavefunction $\Psi(r)=|\Psi(r)|\, exp(i\phi(r))$. As a consequence Cooperpairs are strongly correlated and coherent: they can not move independently and their phases ϕ are strongly coupled. They are said to form a condensate, with all Cooperpairs having exactly the same energy E_F. Clearly, these integer spin particles no longer obey Fermi Dirac statistics, but found to be bosons.

The formation of this condensed state at E_F occupied by the electronpairs reduces the density of states (DOS) of the normal single electrons to zero for electrons near the Fermi energy. Given the finite strenght of the electron-phonon interaction, a finite range of energy around E_F will be affected. In the energyspectrum of the normal electrons this results in a gap symmetrically around the Fermi energy, with a width denoted by 2Δ. The size of the gap depends on temperature and magnetic field. Reducing T further below T_c increases the size of the gap, reaching its maximum at T=0K. Increasing the magnetic field will reduce the gap and at a (material dependent) value H_c it drops to zero, i.e. the superconducting state is suppressed completely. For many superconductors the width of the gap at zero temperature and magnetic field $\Delta(0)$ and the critical temperature T_c are related by $2\Delta(0)\sim 3.5 k_B T_c$, with k_B being Boltzmanns constant. Typical gaps are ~<1 meV with associated T_c ~<10K. Note that in general $2\Delta << E_F$ (semiconductors: E_F~10-100 meV; metals: E_F~5 eV). As our discussion will be largely limited to T=0, we will denote the gap at T=0 simply by 2Δ.
Fig. 6.1 summarises the results from the preceeding discussion.

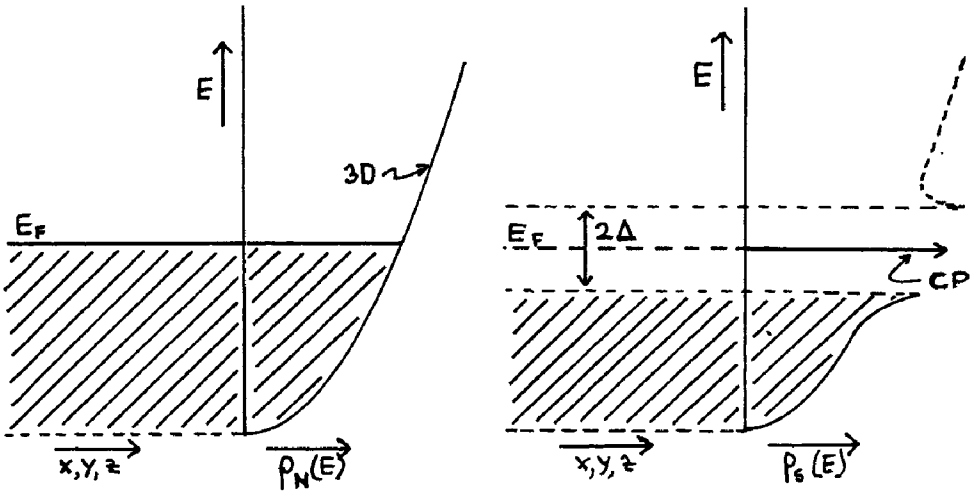


Figure 6.1. The energy spectrum and the density-of-states $\rho(E)$ of electrons; a. in a normal metal or a semiconductor, N; b. in a superconductor, S; CP denotes the condensate of Cooper pairs at the Fermi energy.

6.2 Superconductor-Normal (S-N) interfaces and Andreev reflection.

The question arises how the transport of electrons will be affected if a superconductor (S) will be coupled to a normal (N) material. Fig. 6.2 shows the resulting configuration.

Evidently electrons from the normal region N impinging on the N-S interface with anenergy $E=\varepsilon+E_F$ will experience the change in the density of single electron states in the superconductor S. In particular electrons with an energy $|\varepsilon|<\Delta$ will find no states at all

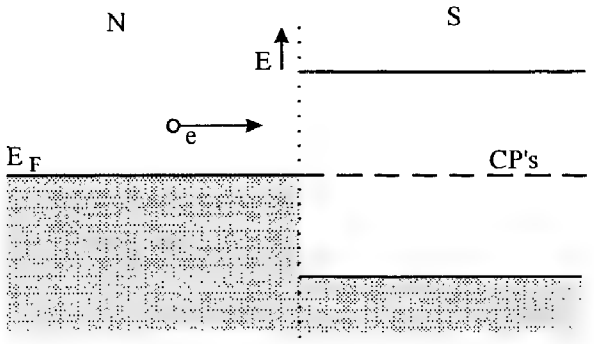


Figure 6.2. The energy spectrum of a superconductor connected to a normal conductor.

available within the superconductor and so at first sight no transport can take place.

In 1964 Andreev studying this system showed that transport can occur assuming a more complicated transmission process which involves 4 particles. An electron in N (Figure 6.2) with energy E_F (i.e. $\varepsilon=0$) incident on the interface enters S, drawing a second electron of opposite momentum (and spin) from the N side. The removal of this second electron from N leaves a hole (with positive charge and negative effective mass) in N at the interface. This hole being fully complementary to the second electron, will have its momentum vector opposite to this (second) electron, and so in the same direction as the k-vector of the initial electron. As the hole has a negative effective mass, its velocity will be opposite to that of the incoming electron, and so it will start to *trace back the path of this electron: it is retroreflected.* Figure 6.3 shows the process in more detail.

Assume the incoming probe electron has a positive energy $\varepsilon=E_{e,1}-E_F<\Delta$ and a momentum $k_e=k_{e,N,1}$. During the Andreev reflection proces at the interface three quantities need to be conserved: energy, charge and momentum.

= The sum of the electron energies $E_{e,1}$ and $E_{e,2}$, and the hole energy E_h before the AR event has to be equal to the sum of the energy of the Cooper pair $2E_F$ and the hole E_h:

$$E_e \equiv E_{e,1} = E_F + \varepsilon = \frac{\hbar^2 k_{e,1}^2}{2|m_e|} \qquad (6.1a)$$

$$E_h = E_F - \varepsilon = \frac{\hbar^2 k_h^2}{2|m_h|} = E_{e,2} = \frac{\hbar^2 k_{e,2}^2}{2|m_e|} \qquad (6.1b)$$

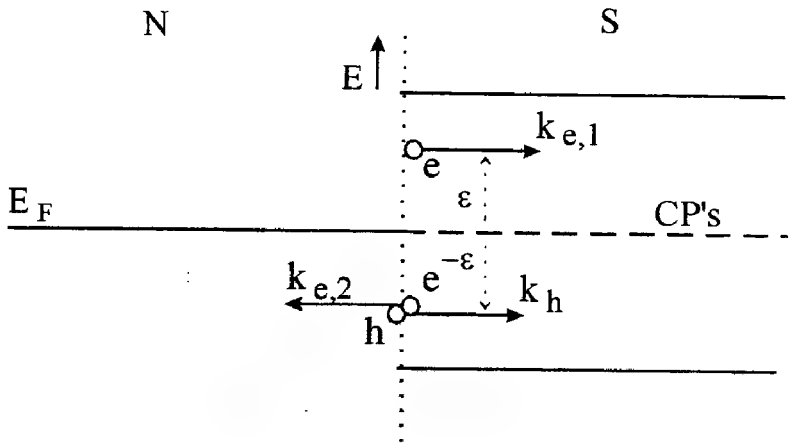


Figure 6.3. The proces of Andreev reflection at the interface of a normal metal N and a superconductor S. The incident electron e at energy ε relative to the Fermi energy is converted into a (retroreflected) hole h and a Cooper pair CP.

and $E_{e,1} + E_{e,2} + E_h = 3E_F - \varepsilon = E_{CP} + E_h$ (6.1c)

Note that the energy of electron 1 is just as much above the Fermi energy as the hole energy is below it.

= The charge of the incident electron 1, $-e$, has to match the sum of the charges of the reflected hole, e, and the Cooper pair charge $2(-e)$:

$$-e = e + 2(-e) \qquad (6.2)$$

= The Cooper pair formed in S combines the incoming electron 1 with wavevector $k_{e,N,1}$ (= $k_{e,S,1}$ in S) with a second electron 2 from N. To allow the formation of the CP this second electron should have a wavevector *opposite* to $k_{e,N,1}$, i.e. $k_{e,S,2} = -k_{e,N,1}$. As the hole in N immediately results from the extraction of the second electron from N into S, the sum of their momenta should be identical zero, or $k_h = -k_{e,S,2}$.
Now we have to remind ourselves that the conservation of energy made $E_e = E_h + 2\varepsilon$, with $0 < \varepsilon < \Delta << E_F$, and so eq. (6.1) yields $|k_{e,N,1}| > |k_h|$ for electron 1 and the reflected hole in N. So, effectively the reflected hole and the incident electron will have *nearly equal* momenta (both in size *and direction*), with the small difference determined by the excess energy ε. In summary:

$$k_h = -k_{e,N,2} = -k_{e,S,2} \leq k_{e,S,1} = k_{e,N,1} \equiv k_e \qquad (6.3)$$

The difference in momentum $\Delta k = k_e - k_h$ is taken up by the Cooper pair. From the momenta we immediatedly can derive the velocities. Taking into account that the effective mass of

the hole is negative, i.e. $m_h = -m_e$, we obtain for the velocities of the two particles in the N region:

$$v_h \cong -v_e \qquad (6.4)$$

In figure 6.4 the process of Andreev reflection is shown in real space, including the velocities of the various particles involved. It again stresses the peculiar retroreflective backtracing of the hole along the path of the original incoming electron. It should be kept in mind that this backtracing is only exact if $\varepsilon=0$ or $\Delta k = k_e - k_h = 0$. This difference in

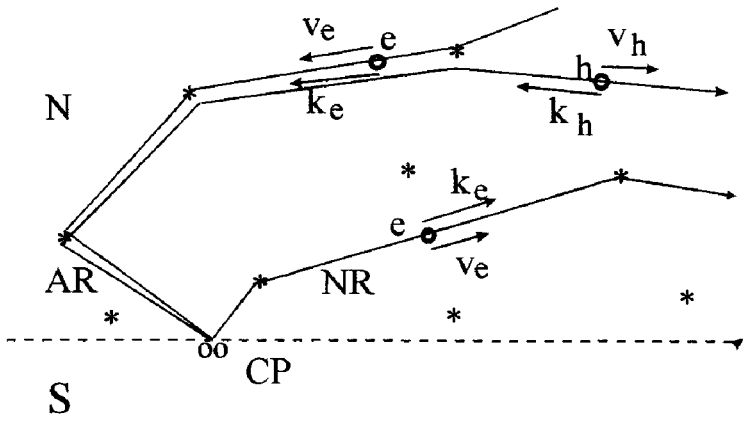


Figure 6.4. Normal and Andreev reflection at an N-S interface. Note the opposite direction of the velocity and momentum vector of the retroreflected hole. The small difference in the paths of the incident electron and Andreev reflected hole is a consequence of the difference in k-vector following eq. (6.5).

momenta Δk can be expressed in the energy difference ε of the electron relative to the Fermi energy:

$$\Delta k \cong \Delta E / (\frac{\hbar^2}{m} k_F)$$

As the energy difference between the electron and hole state $\Delta E = 2\varepsilon$, it immediately follows that

$$\Delta k \cong \frac{2\varepsilon m}{\hbar^2 k_F} = \frac{2\varepsilon}{\hbar v_F} \qquad (6.5)$$

From this expression we can derive over what distance the incoming electron and the retroreflected hole maintain their relative phase. If this difference in phase becomes larger than $\Delta\phi = \pi$ the initial in-phase condition is changed into out-of-phase. This will happen after travelling a distance

$$L = \frac{\pi}{\Delta k} = \frac{h v_{F,N}}{4\varepsilon}$$

For the maximum excess energy $\varepsilon=\Delta$ where Andreev reflection occurs, a new lengthscale ξ_N can be defined by

$$\xi_{N,C} = \frac{h v_{F,N}}{4\Delta} \qquad (6.6a)$$

This length is called the phasecoherence length, in this particular case defined for the case of ballistic transport in the N region, i.e. the "clean" case. To evaluate the coherence length in the diffusive case, we have to realise that the time associated with the dephasing process is given by $\tau_N = \xi_{N,C} / v_F = h / 4\Delta$. From this we immediately obtain the phasecoherence length for the diffusive case (D is the diffusion coefficient, see chapter 4)

$$\xi_{N,D} \equiv \sqrt{D\tau_N} \approx \sqrt{\frac{h}{4\Delta} v_F l_e} = \sqrt{\xi_{N,C} l_e} \qquad (6.6b)$$

From the description of the Andreev process it is clear that the incoming electron effectively "drags" a second electron charge into the superconductor. This means that the total charge traversing the plane of the interface is twice as large as that of the incident electron alone, i.e. the current is effectively doubled. So ideal Andreev reflection doubles the conductance of an N-S system compared to the case that the S-part is in the normal state (e.g. by applying a magnetic field).

The description given above assumes that at the N-S interface ideal Andreev reflection occurs: the incident electron is totally converted into a hole and vice versa. If the interface shows a finite transmission probability T for the incoming particle a fraction of the wave will be reflected maintaining the original particle character, i.e. electron as electron and hole as hole. This normal reflection of course will reduce the fraction of the wave being Andreev reflected. As a consequence the conductance with S in the superconducting state will be less than twice the conductance with S in the normal state. This can be described more precisely. If the transmission probability of the n-th mode is given by T_n, then we know that, with both metals in the normal state, the conductance G_{NN} for M modes being transmitted is given by the common Landauer expression

$$G_{NN} = \frac{2e^2}{h} \sum_{n=1}^{M} T_n \qquad (6.7a)$$

If superconductivity is not suppressed the resulting conductance G_{NS} can be calculated. A rather complicated calculation yields for the conductance

$$G_{NS} = \frac{4e^2}{h} \sum_{n=1}^{M} \frac{T_n^2}{(2 - T_n)^2} \qquad (6.7b)$$

From this expression we can see two things. First, if all modes have unity transmission probability T_n, we immediately see that $G_{NS}=2G_{NN}$, i.e. the doubling mentioned before. Figure 6.5 shows the case for a ballistic point contact system with all $T_n=1$, except the topmost (M-th) mode which is transmitted only partially. Both $2G_{NN}$ (eq. (6.7a)) as well as G_{NS} (eq. 6.7b)) are shown in dependence of the the width of the point contact. Note that the doubling only holds at the plateaux and not in the intermediate transition regions where $T_M<1$.

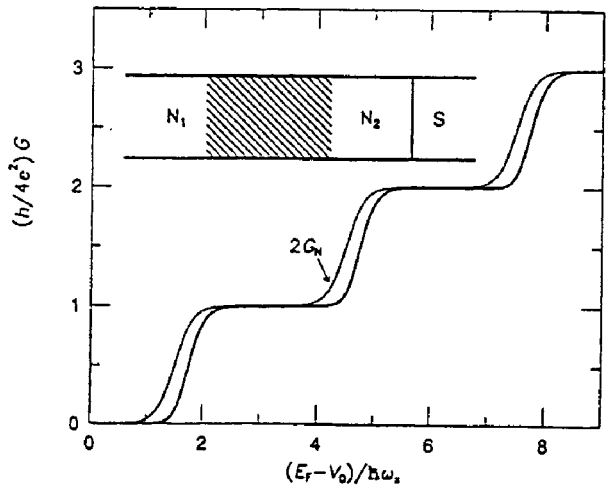


Figure 6.5. The conductance G_{NS} of a point contact at an N-S interface, in dependence of the contact width. G_{NN} is the conductance of the contact if superconductivity is suppressed.

The second case to consider is if the N-S interface is non-ideal, showing only partial transmission. If all modes have $T_n<<1$, eq. (6.7b) shows that $G_{NS} \propto \sum T_n^2$ and so $G_{NS}<<G_{NN}$, i.e. the conductance in the superconducting state is reduced compared to that in the normal state. Note that in this case the conductance goes like the *square* of the transmissions T_n, in contrast to the normal case given by eq. (6.7a). This is an immediate consequence of the *two particles (=electrons)* that have to cross the interface *simultaneously* in the case of Andreev reflection.

In chapter 4 we discussed transport in the diffusive regime in disordered systems. If the phasecoherence length l_ϕ in such system is much larger than the elastic scattering length l_e it was found that the conduction electrons could experience quantum interference, leading to an increased probability for backscattering via closed time-reversed trajectories. This coherent backscattering effect, called weak localisation (WL), manifested itself in a reduction of the conductance. This reduction could be suppressed again by applying a

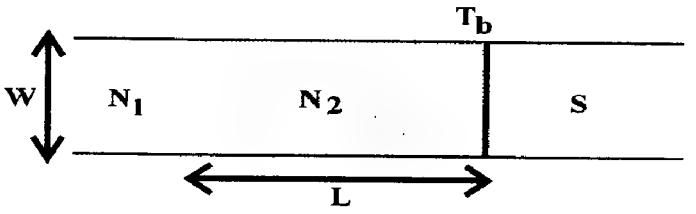


Figure 6.6. A disordered normal conductor N_2 of length $L>> l_e$ and a superconductor S, connected by a barrier with transmission T_b

magnetic field, such that the time-reversal symmetry is broken for the closed trajectories of length $\sim l_\phi$. This field is thus given by

$$B_c \approx \Phi_0 / l_\phi^2 = h / e l_\phi^2 \qquad (6.8)$$

Now we may ask what would happen if a superconductor is put sufficiently close to the disordered normal area. We will show that two distinct phenomena arise, depending on the (single electron) transmission probability of the interface, T_b. Although we will not endeavour to give a formal derivation we intuitively can appreciate what will happen. We assume that phase coherence is preserved up to the N-S interface. Figure 6.6 shows the resulting system. First we will discuss the case of transmission $T_b \sim 1$, followed by the low transmission case $T_b << 1$.

Assume that the length L of the disordered region is smaller than the phase coherence length l_ϕ, and take the transmission at the interface $T_b = 1$, i.e unity Andreev reflection. Consequently each electron travelling along the time-reversed trajectories that lead to WL and impinging on the N-S interface will become accompanied by an Andreev hole (figure 6.7). Take a closed trajectory starting in (and returning to) some position O in N, with part of the trajectory running through S. This trajectory now can be travelled along in fourdifferent ways: by an electron which eithers runs "clockwise (CW)" or "counterclockwise (CCW)", but also by the Andreev hole which also may take it in the two (CW or CCW) directions. At zero excess energy the hole will stay in phase with the electron irrespective of the length of the trajectory (eq. (6.5)). Consequently, in addition to the electron taking the time-reversed CW and CCW paths (leading to the "common" WL) the phasecoherent Andreev hole can follow the same path (also CW and CCW). This will

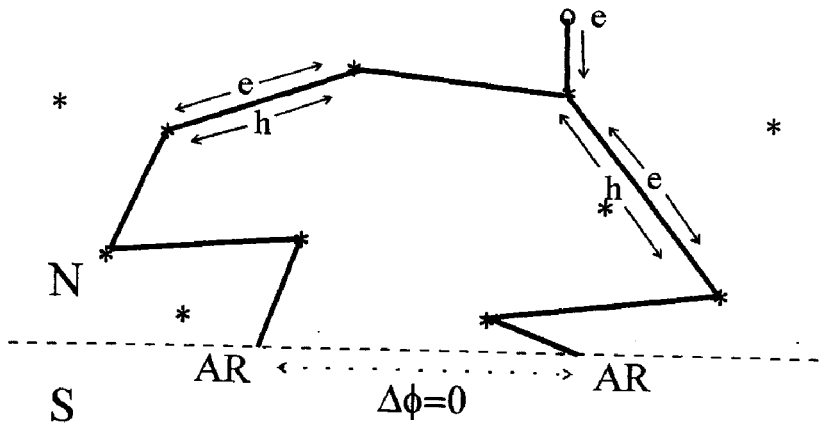


Figure 6.7. Enhanced weak localisation (EWL) in a N-S system, with transport in the normal material in the diffusive regime. The incoming electron can follow the closed trajectory in four ways: CW and CCW as an electron, and CW an CCW as the AR hole.

give an additional contribution to the coherent backscattering of roughly the same size as obtained for the normal state, in this way (approximately) doubling the common weak localisation. This effect is called *Enhanced Weak Localisation* (EWL). Evidently the precise size of the enhancement will depend on the value of the transmission at the N-S interface T_b, the the ratio l_e/L and the number of transmitted channels (or the width W).

Figure 6.8 shows the results of a numerical simulation performed by Beenakker et.al. The

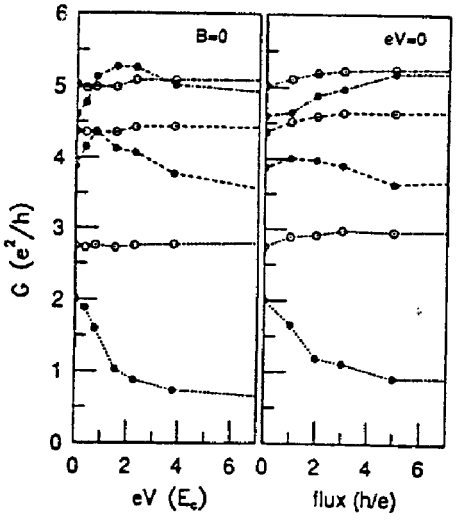


Figure 6.8. G_{NN} (filled dots) and G_{NS} (open dots) of a disordered normal area connected to a superconductor via a barrier of transmittance T_b, in dependence of bias voltage and magnetic field (or flux). T_b=1 (dashed-dotted), =0.6 (dashed) and =0.2 (dotted).

topmost curves (dash-dotted) show the results for T_b=1, both in dependence of a finite bias voltage (left panel) as well as for an applied magnetic field (right panel). For both conductors in the normal state (open dots) the conductance G_{NN} shows (right panel) the characteristic reduction at zero magnetic field due to WL; this is suppressed by a finite flux, visible as a (small) rise of G_{NN}. If in the superconducting state (filled dots) the reduction of the conductance G_{NS} is considerably larger, demonstrating EWL. Note that the *enhanced* fraction of the WL can be suppressed by *both* a finite magnetic field (or flux) as well as a finite bias voltage, in contrast to the "common" WL which only is affected by the field. We can understand this immediately, realising that WL is suppressed by breaking time reversal symmetry. In a normal system only the magnetic field can do so. In the case of Andreev reflection the finite bias leads to a finite difference in energy ε and so in *k*-vector between the electron and Andreev hole (eq. (6.5)). This introduces an additional phaseshift after travelling a certain distance and so it will lead to suppression of the coherent backscattering resulting from the electron-hole interference, i.e. the enhanced part. We can roughly estimate the voltage V_c for suppression of EWL simply as follows. The total *path*lenght travelled by the electron (or hole) during the phasecoherence time $\tau_\phi = l_\phi^2/D$ is given by $L_{path} = \tau_\phi v_F$. With $\varepsilon = eV_c$, eq. (6.5) for Δk, and the condition for suppression $L_{path} . \Delta k \sim \pi$, we find

$$eV_c \approx \frac{\pi \hbar D}{2l_\phi^2} \qquad (6.9)$$

A more accurate calculation shows that the factor of 2 in the denominator should be deleted.

Now let us proceed to the case that the transmission through the barrier is relatively small (T_b<<1). The small transmission means that an electron incident on the N-S interface will have a rather large probability to become reflected as an ordinary electron, reentering the N area (figure 6.9). As $L>>l_e$, the reflected particle will experience many elastic

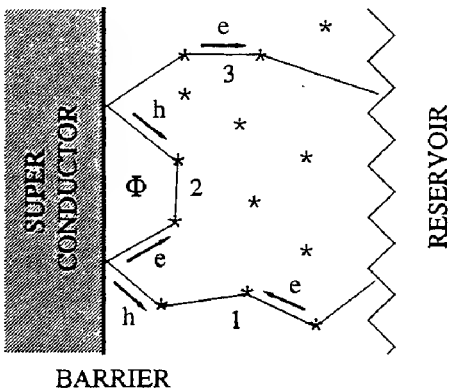


Figure 6.9. Semi-classical picture of multiple Andreev reflection at the N-S interface induced by the strong elastic scattering in the normal area, leading to reflectionless tunnelling.

scattering events in N and so it will have a considerable chance to return to the interface. Figure 6.9 gives a semi-classical picture of this effect, due to van Wees (Phys. Rev. Lett. 69, 510 (1992)). An incoming electron travelling along trajectory 1 will be reflected at *a*. The major fraction of the electron wave will continue as an electron along path 2, but a small fraction will become Andreev reflected and will travel backwards along path 1 as a hole. After a number of elastic scatterings in N along path 2 (part of the) the electron(-wave) will again approach the N-S interface and become reflected at *b*. A (large) electron part will continue along path 3, but an AR hole part will travel back along path 2. Once arriving at *a* this hole wave will be reflected *in the normal way*, and continue along the initial path 1. Note that the "total" hole wave equals the sum of the direct contribution from the AR at *a* and the AR part created at *b*. If the incoming electron has zero excess energy, the hole and electron will have equal wavevector. This implies that the *phase increment* of the electronwave along path 2 exactly cancels the phase *decrement* of the hole along this path, or the two hole contributions returning along trajectory 1 are exactly in phase. In this way the strong elastic scattering in N, leading to multiple impingement of the particle on the interface leads to an *increase* of the probability for Andreev reflection. This increase will evidently lead to an increase of the conductance G_{NS} beyond the "classical" value. This effect is called *Reflectionless Tunnelling* (RT), because it seems as if the dependence of the conductance on the *square* of the transmission T (see eq. (6.7b) and the discussion immediately following it) is violated, or, stated differently, the hole seems to traverse the interface with approximately unity probability.
Note that at zero magnetic field and bias voltage (between the reservoir and S) *all* hole waves returning to the first point of incidence, *a*, will be *in phase*. Evidently, this phase can be affected by the magnetic field (shown as the flux Φ penetrating the closed trajectory 2 with the superconductor S acting as a "phase short circuit") as usual, yielding approximately the same "critical" field B_c as before (eq. 6.8). In addition, just as for the preceeding case of EWL, also here a finite bias voltage will break the phase conjugation between electron and hole, given by eq. (6.9).
In figure 6.8 also the results for reflectionless tunneling are shown obtained from a more realistic, fully quantummechanical calculation. The lower traces (dotted lines) are obtained for a (electron) tunnelling probability T_b=0.2. The open dots (for the normal state conductance G_{NN}) only show the common WL, which is suppressed at a finite flux (see right panel). The filled dots for the superconducting case G_{NS} clearly show the strong increase of the conductance, at zero bias and magnetic field.

Before continuing with more complicated NS systems we will discuss two very recent experiments, which show enhanced weak localisation and reflectionless tunneling. Figure 6.10 shows results obtained on a Nb-InGaAs S-N system, i.e. the normal conductor is a semiconductor (A. Kastalsky et.al., Phys. Rev. Lett. 67, 3026 (1991)). A crucial requirement for all these experiments is that the contact should have a large transmission coefficient T_b. To establish this, a meticulous cleaning of the (InGaAs) surface by etching is essential before evaporating the superconducting metal (Nb). The normalised differential conductance $(1/G_N)dI/dV$, measured in dependence of the bias voltage V, shows two features. At all temperatures below the critical temperature of the superconductor Nb

(T_c=9.2 K) a broad dip develops, resulting from the reduction of the density-of-states of single particles at the Fermi energy. If the temperature is reduced further to below ~1.5 K

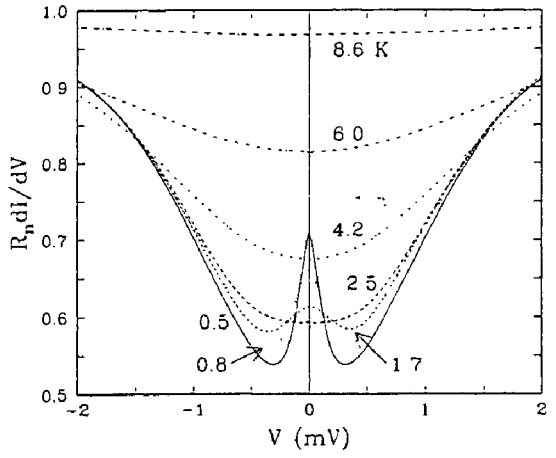


Figure 6.10. Reflectionless tunneling in an N-S system comprising Nb as the superconductor and InGaAs as the N region. The strong increase of the conductance at low temperatures (T<1.5 K) and at zero bias is a clear demonstration of the effect.

a peak starts to emerge. This increase of the conductance at small bias voltages (<~ 0.3 mV) is a manifestation of the effect of reflectionless tunneling.

The second experiment of quantum interference in disordered N-S systems is presented in figure 6.11. In this case the contact is made of Sn, diffused into a GaAs-AlGaAs heterostructure containing a 2DEG. The diffusion process leads to strong local doping by the Sn atoms of valence 4 replacing the Ga atoms of valence 3. The heavely doped,

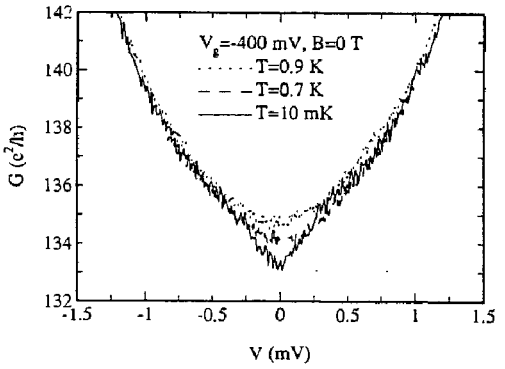


Figure 6.11. Enhanced weak localisation found in a Sn contact made to a 2DEG in a GaAs-AlGaAs heterostructure. The solid curve shows the conductance dip at 10 mK and zero magnetic field. At higher temperature and larger bias voltage (>0.3mV) it is suppressed.

disordered region forms the actual contact between the superconductor and the 2DEG. The system thus formed largely resembles the model as given in figure 6.6. In figure 6.11

the differential conductance (in units of e^2/h) is shown in dependence of bias voltage, at zero magnetic field and at three different temperatures. Apart form the "classical" dip, also seen in the preceding experiment, at temperatures $T<\sim 0.7$ K an additional peakshaped reduction of the conductance starts to develop. This feature is suppressed at bias voltages $> \sim 0.3$ mV, a value similar to that found in the first experiment. It is a manisfestation of the effect of enhanced weak localisation. Note that the enhanced fraction in the conductance approximately equals $2e^2/h$.

6.3 S-N-S systems; multiple coherent Andreev reflection

Now that we have understood the process of Andreev reflection it is interesting to see what effect the introduction of a second S-N interface has on the transport properties. For the purpose of simplicity we will limit ourselves largely to the quasi-1D ballistic case. We will show that, provided the distance between the two interfaces is sufficiently small, a dissipationless (= super-)current can flow between the two superconductors through the N region, the value of which is determined by the phasedifference between the two superconductors.

Fig. 6.12 shows the position-energy diagram of the system we want to discuss. Taking the two superconductors S_1 and S_2 to be the same, we assume that the macroscopic wavefunctions describing the condensates have phases ϕ_1 and ϕ_2 respectively. An electron (e) at an excess energy $\varepsilon=E-E_F$ and travelling to the right will experience Andreev reflection at the rightmost N-S_2 interface, converting into a Cooperpair (CP_2) in superconductor S_2 and a reflected hole (h) which will travel at an energy $-\varepsilon$ backwards towards S_1. This hole, incident on S_1, will be backconverted into an electron by AR, simultanouesly breaking up Cooperpair CP_1 in S_1. The electron, travelling again at an energy $+\varepsilon$, will interfere with the original electron. This whole process has two consequences. First, it results in the transfer of one CP from S1 to S2. Secondly this multiple AR process constitutes a periodic motion. Quantummechanically this motion will be quantised, i.e. it will form bound states with associated eigen energies ε_M. Let us look to these states in more detail.

A state will exist if the electron, after performing one e ->h ->e cycle, returns to its "starting position" with the same phase + $M*2\pi$. To calculate the total phase difference

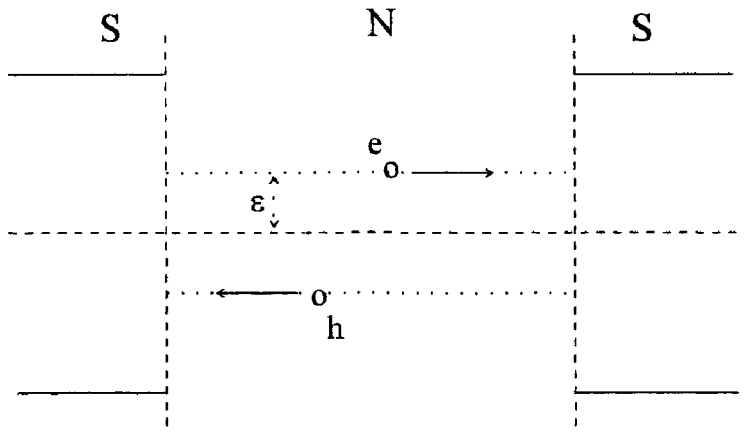


Figure 6.12. An S-N-S system with Andreev reflection taking place at both interfaces. The resulting coherent multiple electron-hole-electron transformation leads to a bound state in the normal region. The state has its electron part above the Fermi energy and its hole part below.

the electron acquires in one roundtrip we have to take into account that at the S-N interface the Andreev reflected particle receives an extra phaseshift. This is given by

$$\delta\phi = \pm\phi_S - \arccos(\varepsilon/\Delta) \qquad (6.10)$$

where ϕ_S equals the phase of the superconductor. Although we will not derive this statement, it is an immediate consequence of the required continuity of the wavefunction and its derivative at the interface. The sign of ϕ_S depends on the nature of the particle, i.e. + for an (incoming) electron and - for a hole.

Take L to denote the length of the normal region between the two superconductors S_1 and S_2. If we take $\varepsilon << \Delta$, the *arccos*-term can be approximated by $-\pi$, and we can write down the condition for an eigenstate directly, yielding

$$k_e^{\pm}L - k_h^{\pm}L \pm (\phi_1 - \phi_2) - \pi = M.2\pi \qquad (6.11a)$$

with M=0, 1, 2, ... The +/- sign distinguishes between states of electrons moving to the left or the right. Rewriting this expression and defining $\phi=\phi_1-\phi_2$, leads to

$$k_e^{\pm} - k_h^{\pm} = \Delta^{\pm}k = \frac{2\varepsilon_M^{\pm}}{\hbar v_F} = \frac{(M+1/2).2\pi \pm \phi}{2L} \qquad (6.11b)$$

So, Δk and ε_M depend on $\Delta\phi$ (including its sign) as well as on the quantum number M. If one takes into account the finite size of the gap relative to the energies of the confined states, than eqs (6.11) have to be modified, and (the second half of) (6.11b) reads

$$\frac{2\varepsilon_M^{\pm}}{\hbar v_F} = \frac{M.2\pi \pm \phi + 2\arccos(\varepsilon_M^{\pm}/\Delta)}{2L} \qquad (6.11c)$$

Fig 6.13 shows these eigenstates ε_M, at zero and at a finite phasedifference ϕ.

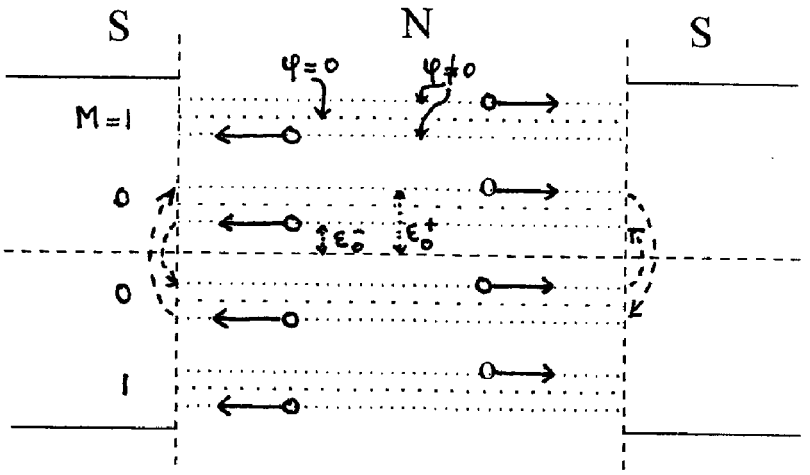


Figure 6.13. Bound states in the N region. The dotted lines are the (degenerate) states at zero phasedifference. The dashed lines indicate the states at non-zero ϕ with the arrows denoting the left- and rightgoing states for electrons ($E>E_F$) and holes ($E<E_F$).

Note that for a given state M the momentum k_e^+ of the right moving electron differs from k_e^- of the left moving electron by an amount that is linearly dependent on the phasedifference between the two superconductors ϕ. This implies that the current carried by the + branch and the - branch will be different as well, and so each state will contribute a net current. Given the similarity in the dependence of the energy versus the phase for all states M, also the total net current will depend linearly on the phasedifference.
To estimate the total current we first have to evaluate the current carried per state M (either + or -). This simply is given by the number of electrons per state, g_M, divided by the time it takes for one electron to make one roundtrip, i.e.

$$I_M^{\pm} = g_M e / \tau_{round} = g_M e \frac{v_F}{2L} \qquad (6.12)$$

From the Pauli principle it is clear that the number of electrons per state simply equals the spin degeneracy factor: $g_M=g_s=2$.
Assume that the lenght L of the normal area is relatively large (i.e. $L>>\xi_N$). From eq. (6.9b) we immediately can deduce that

$$\varepsilon_{M+1}^{\pm} - \varepsilon_M^{\pm} \cong (1/2)\varepsilon_0^{\pm}(\varphi = 0) \qquad (6.13)$$

for the low lying states (M is small). On increasing the phase to $\phi=\pi$ equation (6.9b) also shows that the lowest lying left moving electron state will attain an eigenenergy ε_0^- $(\phi=\pi)=0$. This implies that on a further increase of the phase the left moving electron state should enter the filled Fermi sea, a region only accessable for hole states (see figure 6.13), at the same moment requiring the associated right moving hole state to enter the filled *hole* sea above the Fermi energy. Evidently this is not possible, and the system responds to the phasedifference $\phi>\pi$ by establishing a *right moving electron state above* the Fermi energy and the (AR) associated *left moving hole state* below it. This has two major

consequences. First, because of the change of direction of the M=0 state the current transported by it (eq. (6.12)) will jump from $-ev_F/L$ to $+ev_F/L$, i.e. by an amount

$$\delta I_0 \cong 2ev_F / L \qquad (6.14)$$

As all other states will just continue their decent (- states) or increase (+ states) the *total* current will jump by the same amount. Secondly from figure 6.13 we immediately can see that *all states* at $\phi>\pi$ can be renumbered as

$$\varepsilon_M^+(\phi)=\varepsilon_{M+1}^+(\phi-2\pi)$$

and (6.15)

$$\varepsilon_M^-(\phi)=\varepsilon_{M-1}^-(\phi-2\pi)$$

i.e. as states with a different quantum number M but with a phase $-\pi<\phi<\pi$. This shows that the net current that flows is periodic in the phase with periodicity 2π.
From this discussion we have now reached the following conclusion. Due to the phasecoherent multiple Andreev reflection in the S-N-S system a *net current flows between the superconductors, with a magnitude and sign determined by the phasedifference between the two superconductors*. This implies that the established current is a *supercurrent*, with a maximum or *critical* value given by eq. (6.14), i.e.

$$I_c(0) \cong ev_F / L \qquad (6.16a)$$

Equation (6.16a) is derived assuming the length of the normal part to be long. Reducing L to less than the phasecoherence lenght of the normal material ξ_N will only leave a single bound state in the system (M=0).

Problem: check this statement, using equation (6.9c)

If we insert the value for the ballistic phasecoherence length (eq. (6.6a)) into (6.16a) we directly obtain the critical current for the short junction as

$$I_c(0) \cong \frac{2e\Delta}{\pi\hbar} \qquad (6.16b)$$

Equations (6.16a and b) thus give the value for the supercurrent for a long and a short 1D S-N-S junction respectively, at zero temperature. Including the effect of a non zero temperature on the supercurrent can be approximately done as follows. Once $kT \sim \delta\varepsilon^{+-} = \varepsilon_M^+ - \varepsilon_M^-$, the + and - brach will become mixed. As the difference between such two states determines the net current, the mixing will lead to a reduction of the critical current approximately like

$$I_c(T) = I_c(0)\exp(-\frac{kT}{\delta\varepsilon^{+-}(\varphi=\pi)}) = I_c(0)\exp(-\frac{L}{\xi(T)}) \qquad (6.17)$$

with $\xi(T) = \hbar v_F / 2kT$, denoting the thermal length. In addition to the reduction of the critical current the jump of the supercurrent at $\phi=\pi$ will become smeared out, as the mixing occurs over an energy interval of the order of a few times kT, i.e. it will be effective in a range of phase differences around $\phi=\pi$.
Figure 6.14 summarises the general behaviour of the supercurrent as a function of the phasedifference.

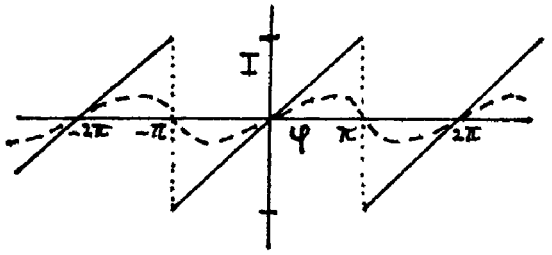


Figure 6.14. The supercurrent in an S-N-S system in dependence of the difference in phase between the two superconductors, at T=0 (solid line) and at a non-zero temperature (dashed line).

Note that the smearing of the current transients around $\phi=(2K+1)\pi$ with K=0, +-1, +-2, .., results in an approximately sinusoidal behaviour of the supercurrent, $I(\phi)\cong I_c \sin\phi$.

All results obtained are limited to the single channel 1D case in the ballistic regime. If more channels are present, say N_c, we simply can add the net currents for each channel. As each channel contributes the same amount of supercurrent we find (for the case of a short junction)

$$I_c(0) \cong N_c \frac{2e\Delta}{\pi\hbar} \qquad (6.18)$$

or the *supercurrent is quantised*, with its value directly determined by the the *number of conducting channels*. Remember that this number also determines the conductance of the system if it is in the normal state, as given by the common Landauer expression, $G_n=N_c.2e^2/h$. So, in case the N region is a quantum point contact, the conductance in the normal state and the supercurrent in the superconducting state should behave in a similar way. Note that, in contrast to the universal value of the steps $2e^2/h$ in the quantised conductance, no such universality holds for the size of the steps in the supercurrent.

Combining eq. (6.18) for the multichannel critical current of a short junction with the normal state resistance $R_n=1/G_n$ yields for the product

$$I_c R_n \cong \frac{2\Delta}{e} \qquad (6.19)$$

Before closing the chapter we want to discuss some experimental aspects. S-N-S systems have been investigated already for a long time, with a strong emphasis on conventional metals taken for the N region. In many experiments supercurrents were found and studied in dependence of temperature and magnetic field as well as the thickness of the normal layer, both in the ballistic and the diffusive regime. Fewer experiments have been done with a semiconductor as the N material. In the majority of these cases the N region was in the diffusive limit, a regime that we have not discussed above.
Up to this moment no experimental result exists that firmly confirms the prediction of the quantisation of the supercurrent as given by eq. (6.18). There is however a recent

experiment in a fully metallic system that points in the correct direction. Figure 6.15 shows the main parts of the set up of this so called break junction experiment (C.J. Muller et.al, Phys. Rev. Lett. 69, 140 (1992)). Employing the elastic bending of a strip, the

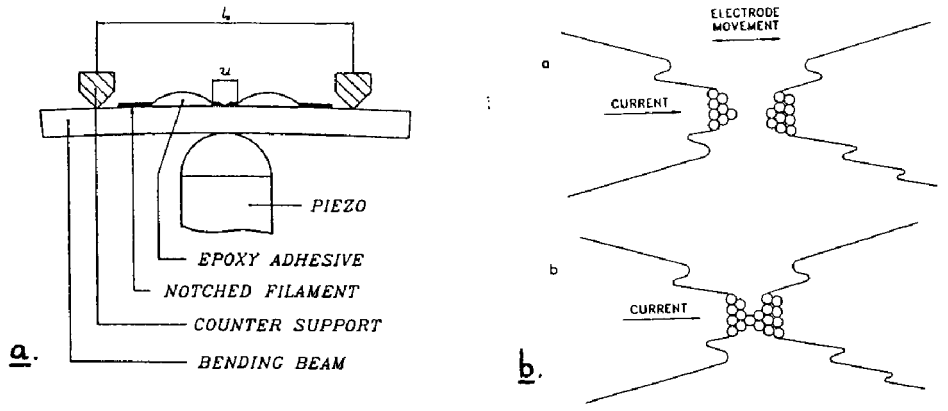


Figure 6.15. A break junction experiment. a.shows the experimental set-up. The thin wire (or filament) with a weak central spot is glued onto the bending beam. By applying a voltage to the piezo elements the amount of bending can be controlled. In this way (b.) the weak spot breaks and the separation of the two sides of the (broken) wire can be varied.

end points of two wires glued onto the strip can be made to have a controllably variable contact. The bending of the strip is obtained by employing a piezo element as a pushing rod against the elastic strip. The gap between the two wires can be controlled to a *pm* ($=10^{-12}$m) scale, i.e. the contact can be manipulated on the scale of the size of single atoms. As a consequence the conductance, determined by the number of contacting atoms as the number of channels N_c, may increase stepwise. In the most ideal case each step could correspond to a single atom subsequently (dis-)connecting the wire ends, and so the step in the conductance should be $2e^2/h$.
Figure 6.16 shows the experimental results obtained for a Nb wire. The lowest panel shows the voltage driving the piezo element that controls the bending state of the strip, and so the separation of the wire ends. As anticipated, a larger voltage resulting in an increase in contact gap, leads to an increase of the resistance with the wire in the normal state (second panel from the top). The typical steps in the resistance of ΔR~20-60Ω at R~600 Ω are equivalent to steps in the conductance $|\Delta G|=|\Delta R| /R^2$~$(2\text{-}6)e^2/h$, i.e. roughly of the right order of magnitude. Note that the resistance varies ~ 50% over the bending range. With the system in the superconducting state the supercurrent is measured, the results of which are shown in the figure in the third panel from the top. Again the current

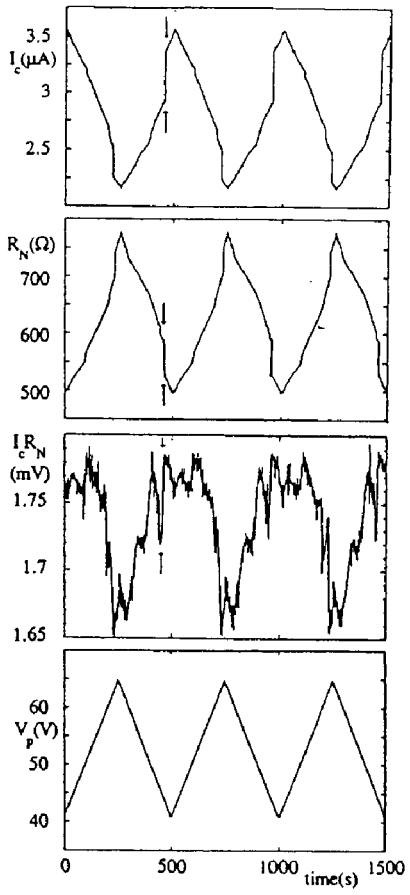


Figure 6.16. Results of a break junction experiment on a broken Nb wire in the normal or superconducting state. The lowest panel shows the (time sweep of the) voltage applied to the piezo that controls the contact gap of the broken wire. The top panel shows the critical current while the second panel presents the resistance of the contact in the normal state. The third panel gives the product of the two.

roughly varies some 50% over the full range. Calculating the I_cR_n product (third panel from the top) shows an average value of ~1.7 mV, with variations of only ~7% total, i.e. much less than those of the current or resistance. Note that the product approximately agrees with the value of the gap of niobium of ~1.5meV.

It should be emphasised that these results do not rigidly confirm the prediction of the quantisation of the supercurrent as the spread in resistance and current steps is much too large. Further experiments, of this type and by using quantum point contacts in semiconductor systems, are needed, and work is progressing in this direction.

General references to the subject:

1. "Andreev reflection and the Josephson effect in a quantum point contact", H. v. Houten and C.W.J. Beenakker, Physica B175, 187-197 (1991)

7. Coulomb interaction and charging effects

Key words: capacitance; charging energy; Coulomb blockade; tunnelling; quantum resistance; single electron transport; Josephson junction; Josephson energy; single Cooperpair transport.

7.1. Introduction

It is known for already a few millennia that under certain conditions matter may become charged. In this state the charged piece of material will exercise a force on a nearby object. Near the end of the 19th century it became clear that this force or interaction resulted from the existence of particles carrying a well defined charge, which became known as electrons. In honour of one of the persons who made crucial contributions to our present day understanding of this force and the particles generating it, this mechanism is denoted as Coulomb interaction.
In the preceding chapters we have neglected any influence of the Coulomb interaction between electrons. In metallic systems this approximation is found to be remarkably successful, reflected in the applicability of the free electron model to a broad variety of solid state problems. In this chapter we will study effects resulting from the electrostatic interaction of electrons, both mutually and with an external electrostatic potential. We will see that in small structures that are (nearly) isolated from its leads, the electrostatic energy of a single (additional) electron, -called the *charging energy*-, may lead to a blockage for the transport of electrons through this structure. This so called *Coulomb blockade* manifests itself at temperatures such that the charging energy exceeds the thermal energy (see section 1.2.b). In addition to the role of the ratio of these two energies also the degree of isolation of the structure, or "localisation" of the charge on it, will be found to be crucial for these phenomena.

In section 7.2 we will see how the single electron charging energy affects the electron transport, both through metallic junction-type structures and semiconductor quantum dot structures. It will become clear that employing the charging energy allows the controlled manipulation of single electrons, both to study the effect as such as well as employing it for intriguing applications like charge detection and accurate current generation. In section 7.3 metallic junctions in the superconducting state will be discussed and we will see that the "unit charge" to be considered is that of a Cooperpair, i.e. two times the electron charge.

7.2. Charging effects in normal-state systems

In this section we will introduce the effects resulting from the addition of a single electron onto a small conducting object. In subsection 1.d. of chapter 1 the phenomenon has been discussed in a very brief and rudimentary way, and it is the present aim to underpin it more thoroughly. Notifying the capacitance of the small object or island, we

will see that the energy related to the addition of the smallest unit of charge, -i.e. one electron-, leads to such energy requirements that under certain conditions these can not be fulfilled by the sources available in the circuit. This then leads to the impediment of charge flow through the island: the Coulomb blockade. We concentrate on normal metal (or semiconductor) bodies and how the electron transport through it is behaving.

7.2.a. *Dimensions and capacitances of conducting bodies*

From elementary electrodynamics we know that an isolated conducting body suspended in a dielectric environment (including free space) shows a linear relation between the charge Q residing on it and its potential ϕ relative to infinity (commonly taken to be the zero of potential). The factor of proportionality is the capacitance C of the body, with $C=Q/\phi$. The associated electrostatic energy is simply given by $E=Q^2/2C$. Apart from the dielectric constant of the environment, the capacitance of the body is completely determined by the geometry of the system, i.e. by sizes and distances. For the sake of completeness three examples of the capacitance of commonly encountered configurations are given. For a sphere of diameter d in infinite space with a dielectric constant $\varepsilon=\varepsilon_r\varepsilon_0$ (with ε_0=8.854188*10^{-12} F/m) the capacitance is given by

$$C_{sphere}= \varepsilon\, 2\pi\, d \qquad (7.1a)$$

Squeezing one dimension of the sphere to form a disc with diameter d yields a capacitance (again in infinite space)

$$C_{disc}= \varepsilon\, 4d \qquad (7.1b)$$

If on the other hand the disc of area $S=\pi\, d^2/4$ is brought near to a large plane conductor at zero potential, such that the two planes (disc and conductor) are in parallel at a distance a, the capacitance is (approximately!) given by the usual "flat plate" expression

$$C_{disc\text{-}plate}= \varepsilon\, S/a = \varepsilon\, \pi d^2/4a \qquad (7.1c)$$

It is important to note that reducing the typical dimension d of the isolated body leads to a decrease of the capacitance of the body, i.e. a body of small size has a small capacitance.

7.2.b. *Conditions for single electron charging*

In large bodies containing many (free) conduction electrons, the charge seems to behave as a fluid, acting as a continuous variable. We know, however, that nature only provides charge in quantised units q=-|e|, in this way only providing total charges $Q=Ne$, with N being integer. As a consequence the electrostatic energy due to the addition of the minimum amount of charge -|e| is given by

$$E_C=e^2/2C \qquad (7.2)$$

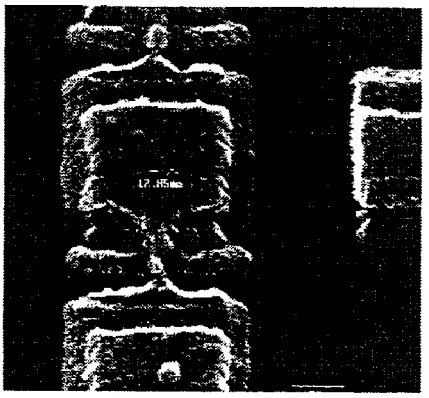
This important energy scale is called the (single electron) *charging energy*. As we have seen that *reducing the dimensions* of a body leads to a decrease of its capacitance, from eq. (7.2) it immediately becomes clear that this also leads to an *increase of the charging energy*. From this it can be anticipated that whenever this energy scale becomes dominant, it will start affect the transport properties through the body. Comparing it to

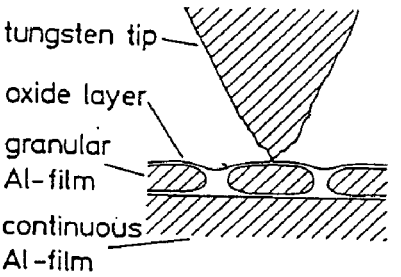
the thermal energy, leads to the *first condition* for charging effects to become effective, given by

$$E_C >> kT \qquad (7.3)$$

This leads to the conclusion that in order to study these phenomena small systems at low temperatures are required.

To investigate the effect of the charging energy we need a structure which allows us to feed current through it. This means that it should be somehow connected to leads. Fig. 7.1 shows three examples of devices which have shown single electron charging effects. Each circuit contains a conducting central section with two leads connected to it. Without providing arguments for it at this moment, it is crucial for the occurrence of charging effects that these leads have a *large resistance*: we will return to this condition shortly. Fig. 7.1a displays a so called metallic double junction device: a central metallic island (e.g. made out of Al) evaporated onto an insulating substrate (e.g. oxidised Si), connected to two metallic leads. Before attaching the leads, the central section is oxidised to form an *extremely thin* (a few monolayers) insulating cover. In this way tunnelbarriers are formed between the leads and the central island. Such a tunnelbarrier forms the high resistance in the leads. A typical size of a tunnelbarrier junction is 100*100 nm^2, showing a capacitance of ~ 10^{-15} F (or 1 fF). The central island with typical dimensions of 1000*100 nm^2 leads to an additional capacitance of some 0.1 fF. Fig 7.1b shows a second example: a Scanning Tunnelling Microscope (STM) tip, pointing to a small metal globule embedded in an insulating matrix on top of a conducting substrate (compare to figure 1.6). Now the tip-to-globule and the globule-to-substrate insulators





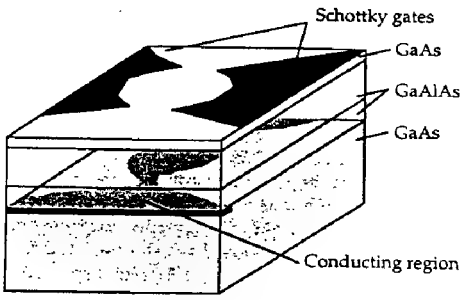


Figure 7.1. Three typical circuits showing single electron charging effects. a. two metal-insulator-metal tunneljunctions in series; b. a STM tip in close proximity to a metallic grain, embedded in a dielectric and placed near a conducting substrate; c. a quantum dot, electrostatically induced in a 2DEG in a semiconductor.

form the tunnelbarriers. With a metallic globule of ~10nm, eq. (7.1a) leads to a typical capacitance $C \sim 10^{-18}$ F. Effectively the total capacitance will be larger due to the relative dielectric constant of the environment and the proximity of the conducting substrate and the STM tip.

Problem: Estimate below what temperature charging effects may become effective with such a 10nm globule.

Fig 7.1c shows a quantum dot made in a GaAs-AlGaAs heterostructure containing a high-mobility 2DEG at the interface between the two constituting semiconductors (see figure 1.13 and Appendix A). From the six gates on the surface which all together form the dot, two pairs are used to form two tunnelbarriers for the high resistance leads; the remaining two gates are employed to complete the dot as such. The capacitance of the system is now mostly determined by that of the central island. With a typical size of 300 nm diameter, eq. (7.1b) yields $C \sim 0.1$ fF, or a charging energy equivalent to $T \sim 10$K.

Now that we have encountered some actual structures, we want to discuss the requirements for the lead resistance, in particular why this should be large. In a piece of bulk metal the free conduction electrons behave nearly like a liquid. This liquid-like state is an immediate consequence of the Bloch states of the electrons, extending far through the metal crystal. The extended nature of these electron states makes it a rather meaningless question to ask how many electrons there are within a small given volume somewhere inside the large piece of bulk metal. This suggests that the amount of charge residing in a given volume is well-defined only *if this volume is rather well separated from its environment*, i.e. it should be connected by leads transmitting the electrons only weakly. To quantify this more precisely, figure 7.2 schematically shows a typical structure as discussed in figure 7.1. The central island with a capacitance C_Σ is connected to two electron reservoirs by resistors R_1 and R_2 respectively. As we have seen the typical energy it takes for a single electron to be put onto the island is the charging energy E_C. In order for the charge at the central island to be well-defined, the typical time for the charge to leak away to any of the two reservoirs should be sufficiently large. More precisely, it is required that the (quantum mechanical) energy uncertainty associated with this leak-out time should be small compared to the charging energy E_C.

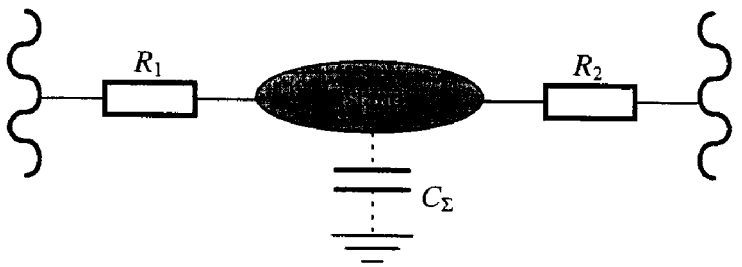


Figure 7.2. Schematic diagram of a single electron charging circuit. The central island, with total capacitance C_Σ is connected to two electron reservoirs by large resistors R_1 and R_2.

The leak-out time is given by the RC time

$$\tau_{RC} = C_\Sigma R_{12} = C_\Sigma / G_{12} \qquad (7.4a)$$

with $G_{12}=G_1+G_2=1/R_1 + 1/R_2$, i.e. the total conductance for electrons to leave the island. The energy uncertainty associated with this time is given by

$$\delta E = \hbar / \tau_{RC} \qquad (7.4b)$$

For the charge to be well-defined and localised at the island we thus should meet the condition

$$\delta E < E_C = e^2 / 2C_\Sigma \qquad (7.5)$$

Combining eq. (7.4) and (7.5) immediately yields a condition for the total "leakage" conductance of the island to the leads

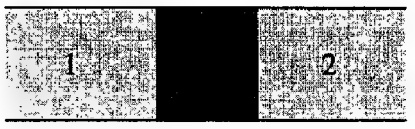
$$G_{12} < e^2 / 2\hbar = \pi e^2 / h \approx 2e^2 / h \qquad (7.6)$$

This is the *second condition* for the occurrence of single electron charging effects. Note that this condition is equivalent to the requirement that the total transmission through the barriers is less than one (see eq. (4.12)). Or, stated differently, the transmission needs to be so small that *quantum fluctuations of the charge* through the tunnelbarrier are effectively suppressed to less than a unit charge |e|.

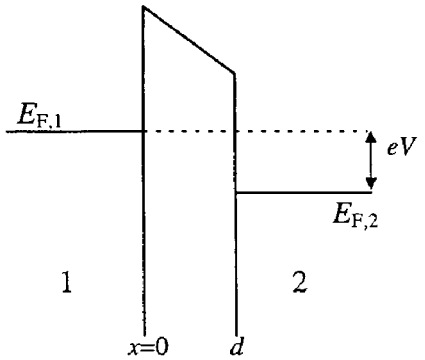
7.2.c. *Tunnelling through a single tunnelbarrier or tunneljunction, and single electron charging effects.*

Now we concentrate on a single tunnelbarrier junction, as shown in figure 7.3. As shown, the barrier is represented by the potential energy rise in the range x=0-d between two conducting regions, i.e. electron reservoirs (e.g. see section 3.2). As the potential energy in the middle of the structure is larger than the maximum kinetic (=Fermi) energy of the electrons in the reservoirs, the electrons are classically not allowed to traverse this barrier region. Penetrating the barrier is only possible via the quantum mechanical mechanism of tunnelling. Let us consider the mechanism of tunnelling in more detail. In doing so we will "derive" a simple expression for the tunnel rate, or the number of tunnel events per second, and show how this depends on the energy difference before and after tunnelling (i.e. between the initial and final state) and on the transmission of the barrier.
As tunnelling is a quantum effect, the tunnelling rate Γ for a single junction is derived starting from the Hamiltonian of the system consisting of the two metal reservoirs and the interconnecting barrier. We assume that the exchange of electrons between the two reservoirs via the barrier is moderate. In quantum mechanical language this means that the coupling between the two reservoirs is weak, and this allows us to describe the total system of two reservoirs and barrier by a Hamiltonian $H=H_1+H_2+H_t$. Here H_1 and H_2 denote the Hamiltonians of the electrons in the reservoirs, taken as if there was no coupling between the two, and H_t describes the exchange of electrons via tunnelling. Writing the Hamiltonian in this way implies that we assume that *perturbation theory* is applicable. In that case we can use the celebrated Fermi Golden Rule to evaluate the transition probability or rate Γ between the i(-nitial) and f(inal) state of the system

$$\Gamma(i \to f) = \frac{2\pi}{\hbar} |< \Psi_i | H_t | \Psi_f >|^2 \, \delta(E_i - E_f) \qquad (7.7)$$



a. insulator



b.

Figure 7.3. A tunnel barrier between two metallic electron reservoirs 1 and 2 (top panel, a.). In the lower panel b. the energy landscape (E-x) is shown, with $E_{F1,2}$ denoting the Fermi energies. A bias voltage V applied between the metals 1 and 2 results in a displacement eV of the Fermi energies at the two sides.

The delta function $\delta(E_i\text{-}E_f)$ enforces energy conservation in the (non dissepative!) tunnelling process. The *current* resulting from this rate is simply given by $I=|e|\Gamma$. The tunnelling matrix element $t_{if}(E)=\langle\Psi_i|H_t|\Psi_f\rangle$ denotes the overlap between the wavefunctions before and after tunnelling (i.e. from reservoir 1 to 2) and so it represents the transmission coefficient of the barrier, at energy $E_i=E_f$.

Before continuing to evaluate the transition probability (7.7) it is very important to understand and appreciate the conditions for the validity of the perturbation approach, and the consequential limitations for the applicability of Fermi's Golden Rule. It assumes the coupling between the two metallic electron reservoirs to be so weak that the states Ψ_i and Ψ_f *themselves* are not affected by it. This also implies that eq. (7.7) only describes the probability for a single transition, from the initial to the final state, without allowing the particle to tunnel back to the initial state. In contrast, if the coupling becomes strong the two separate states restricted to reservoir 1 and 2 respectively merge into states that extends all the way throughout the whole metal-insulator-metal system. This extension implies that the particle can "cross" the "insulator" many times. It is this multi-passing (while preserving its phase: it is a single quantum state!) which makes Fermi's Golden Rule not applicable any more.

Now let us return to the derivation of the tunnel rates. Eq. (7.7) holds for each single pair of states. If we assume that all tunnelling events between all individual initial and final states behave independently (which is by no means obvious!), we can obtain the total rate by simply summing all individual contributions. In a metal the density-of-states ρ is large (see section 1.2.b) and so the sum can be replaced by an integral. So, to calculate the total rate we have to multiply (7.7) by the *number of occupied initial states* in the considered range of energy, i.e. by $\rho_1(E)*f(E)$, with $\rho_1(E)$ being the DOS in electrode 1 and $f(E)$ denoting the Fermi function describing the effect of the temperature on the occupation of states. In exactly the same way also the number of *empty final states* has to be taken into account by multiplying by $\rho_2(E')*(1\text{-}f(E'))$, where the prime ' denotes that the two reservoirs may be displaced in energy relative to each other. In figure 7.3 the displacement eV was taken to result from an externally applied bias voltage. However,

just as well it may result from the energy difference associated with the tunnelling of an electron, and so somehow with the charging energy. We will return to that below. Now we can write the total rate of tunnelling from reservoir 1 to 2 as

$$\Gamma_{12} = \frac{2\pi}{\hbar} \int_{-\infty}^{\infty} |t_{12}(E)|^2 \rho_1(E-E_1) f(E-E_1) \rho_2(E-E_2) \{1-f(E-E_2)\} dE \qquad (7.8)$$

In a normal metal $t(E)$ and $\rho(E)$ are (approximately) independent of energy, provided the energy difference $|E_1-E_2| << E_{Fermi}$. Hence, the integral only depends on the overlap of the two Fermi functions $f(E)$, and so on the energy change $\Delta U=E_1-E_2$ that is acquired in the tunnelling process. Without going through all the (relatively simple) algebraic details, eq. (7.8) can now be rewritten as

$$\Gamma_{12} = \frac{\Delta U}{e^2} G_t \frac{1}{\exp(\Delta U / kT) - 1} \qquad (7.9)$$

Here $G_t=(e^2/h)*\rho_1*\rho_2*|t_{12}|^2$ is the conductance of the tunnelling barrier. Compare this expression to the more simple expression for the ballistic case $G_t=(e^2/h)*N_{ch}$, with N_{ch} being the number of completely transmitted 1D channels (i.e. $|t_{12}|^2=1$). This can be understood as now only individual states in the leads 1 and 2 can "talk to each other", or $\rho_1*\rho_2$ just counts the number of corresponding states in the leads, i.e. it represents N_{ch}. It should be stressed again that eq. (7.9) only holds for *weak coupling* between the electrodes, i.e. $G_t<<e^2/h$. Deviations will also occur if the density-of-states becomes comparable to other energy scales, like kT. We will come back to this aspect in the discussion of semiconductor structures, where the confinement leads to such 0D splittings.

At zero temperature (7.9) can be simplified further to

$$\Gamma = \begin{cases} -\frac{\Delta U}{e^2} G_t & \text{for } \Delta U<0 \\ 0 & \text{for } \Delta U>0 \end{cases} \qquad (7.10)$$

From the preceding expressions we can derive in a quantitative way the currents resulting from tunnel processes through single and multiple junction systems, provided the energy difference and the transmittance of the barrier(s) are known.

In practice two types of barriers can be made. In metal systems an oxide barrier can be fabricated which has a height of typically a few eV (i.e. $>>e/2C$); in this case the transmission and conductance of these barriers is nearly independent for bias voltages in the range determined by the typical charging energies of a few mV. In contrast, in semiconductors a tunnelbarrier induced by gates (see section 4.2) is rather low (~10-100 meV) and broad (tens of nm's), leading to a significant dependence of its transmission on the bias.

Let us now consider the single junction as shown in figure 7.3, of such a size that the energy will be affected by its capacitance. We will discuss two limiting cases. First the junction is connected to a voltage source (of a very low internal resistance), and secondly we will discuss the case that a current source (which has a very large internal resistance) feeds the junction.

We start with the application of a voltage source. If the barrier of resistance $R_t=1/G_t$ has a capacitance C, the application of this bias voltage V or electrochemical potential difference $\Delta\mu=eV$ across the device results in the build up of charges $+Q$ and $-Q$ at the two sides of the barrier.
What values can this charge take, and does it "suffer" from the quantised nature of charge? In the macroscopic metallic leads ending at the barrier the electrons are in extended states, i.e. they can move freely. Consequently the accumulated charges $+Q$ and $-Q$ effectively result from a *shift in the average positions* of the electrons on the two sides of the barrier. This shift can be infinitesimally small and so it leads to a *continuously variable polarisation charge*, which is linearly dependent on the difference in electrochemical potential. In particular it follows the common dependence $Q=CV$. In contrast, with the tunnelresistance $R_t >> h/2e^2$ (i.e. fulfilling our second condition, eq. (7.6)), *any charge that would tunnel* will be well-quantised in units e.
Now we simply can derive the current that flows through the junction. Inserting the energy difference $\Delta U=-|e|V$ into eq. (7.10) one finds $I=|e|\Gamma=VG_t=V/R_t$, which is in agreement with Ohm's law.

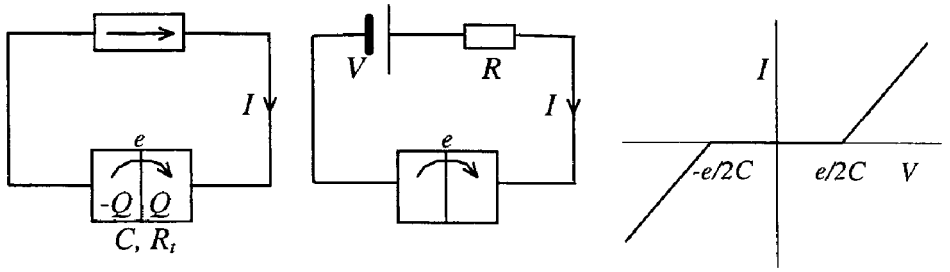
Secondly, in order to study the opposite limit, we connect a current source to the junction (figure 7.4a). This allows us to study the time dependent behaviour of the charge on a single junction. Such a source has an infinitely large internal resistance R_S, and forces a constant current I through the junction, irrespective of the processes occurring in it. If a single electron with a charge $-|e|$ tunnels, the charge changes from Q to $Q-|e|$. The electrostatic energy of the system changes from $U_{initial}$ before tunnelling to U_{final} after the event, with

$$\Delta U \equiv U_{final} - U_{initial} = \frac{(Q-|e|)^2}{2C} - \frac{Q^2}{2C} - = -\frac{|e|}{C}(Q - \frac{|e|}{2}) \qquad (7.11)$$

At $T=0$ tunnelling will occur only if $\Delta U<0$, i.e. whenever the system may enter a lower energy state, and this condition yields

$$Q>|e|/2 \quad \text{or} \quad V=Q/C>|e|/2C \qquad (7.12)$$

If R_t is infinite (i.e., a perfect capacitor) evidently no tunnel events can take place and so



a. b. c.

Figure 7.4. A tunneljunction of capacitance C and tunnel resistance R_t connected to a current source (a.). Panel b. is the circuit equivalent using a voltage source and large series resistance R. In panel c. the typical shape of the resulting I-V characteristic is shown, demonstrating the suppressed current at small bias due to Coulomb blockade.

the constant current would imply a linear increase of the charge Q with time. However if R_t is finite, tunnelling is allowed whenever the condition of eq. (7.12) holds. Fig 7.5 depicts the resulting behaviour of the charge Q versus time t. Note that the total charge Q oscillates between *approximately* $-e/2$ and $+e/2$, and so the voltage between $-e/2C$ and $+e/2C$, i.e. during half of the period of the oscillation the polarity of the voltage over the junction is *reversed* as compared to what one would assume intuitively! In addition the sawtooth oscillatory behaviour of the charge versus time (at a sufficiently small current) makes the *average* voltage across the junction to be approximately zero. The oscillatory behaviour of the charge occurs at an *average* period of $\Delta t=e/I$. This results in *Single Electron Tunnelling (SET)* oscillations at a frequency $f_{SET}=1/\Delta t=I/e$.
The important conclusion is that the Coulomb interaction of the electrons enforces that the electrons flow through the junction is a *time correlated* way: the finite charging energy effectively regulates the traversal of electrons through the system from *completely random* to *highly sequential.*

The condition of eq. (7.12) defines, at zero temperature, the earliest moment in time that tunnelling becomes energetically allowed. The *actual* moment the event will take place is governed by the RC-leakage time of the junction, as this limits the rate at which the charge can be changed. This can be seen easily by inserting eq. (7.11) into eq. (7.12), and taking the a linear increase of the charge with time, following $Q=It|e|$, with I being the externally applied current. This also shows that the probability for tunnelling increases with time once the condition (7.12) is fulfilled. As the leakage of the charge is a classically statistical process (it is governed by the probability just mentioned), consecutive time intervals between tunnelling events will vary. This is indicated in figure 7.5 by the dotted lines. Due to these fluctuations the Fourier spectrum of the SET oscillation will not just show a sharp peak at the SET frequency I/e, but it will be broadened. This becomes particularly important if τ_{RC} is no longer small compared to the SET period Δt, or $I\sim e/RC$.
A non-zero temperature will induce additional fluctuations in the moments in time at which tunnelling occurs. This can be simply seen by introducing an additional energy term in eq. (7.12) which is allowed to fluctuate between $\sim -kT$ and $+kT$. From this it is also evident that by the time this thermal excitation kT becomes comparable to $e^2/2C$ the

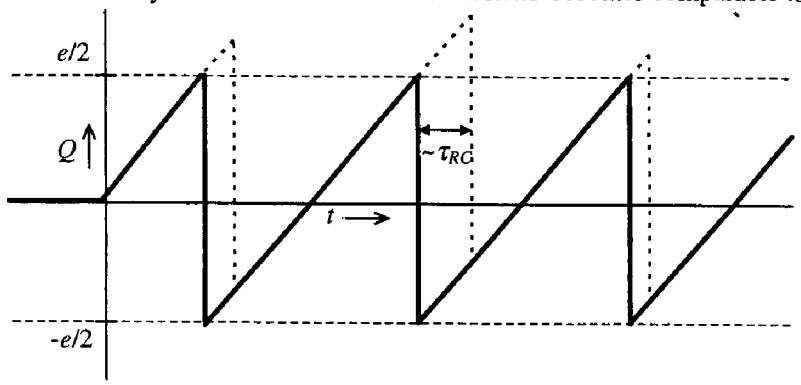


Figure 7.5. The time dependence of the charge on the tunneljunction during SET oscillations. The dashed lines indicate the condition for tunnelling $\Delta U=0$ (eq. 7.12). The actual tunnelling event (dotted line) lags in time in a statistical way on a time scale τ_{RC}.

time correlation in the tunnelling process will be broken and SET oscillations become suppressed.

The current source depicted in figure 7.4a can be schematically represented by a voltage source and a large series resistor R_S (compared to the relevant scale of resistance defined by our second condition $h/2e^2$, eq. (7.6)), figure 7.4b. From eq. (7.12) we can see that, to allow tunnelling, the voltage across the junction has to be larger than $e/2C$. This immediately imposes a lower limit on the voltage V of the source to allow a finite tunnel current to flow. So, only for voltages $|V|>|e/2C$ the current I will be non-zero, while for smaller $|V|$ the current will be zero, as shown schematically in figure 7.4c. This is the famous *Coulomb blockade of single electron tunnelling.*

Now that we have discussed the behaviour of a single junction driven by a constant current, i.e. from a source of very large source resistance, we also want to consider the opposite limit of connecting a voltage source to the junction. The resulting circuit can be envisioned simply by replacing the current source of figure 7.4a by a voltage source. This is equivalent to taking the source resistance in figure 7.4b to be small or $R\sim0$. Evidently, now the voltage across the junction is fixed by the source keeping also the charge constant at $Q=CV$. Because the source rigidly fixes the charge on the tunnel capacitor, any tunnelling charge is *instantaneously* compensated by the voltage source, and so the transport through the junction is only determined by its conductance $G_t=1/R_t$ for all values of the applied voltage V. So, connecting the junction to *a low-resistance source suppresses the Coulomb blockade.*
From these two limiting cases of large and small resistance (or, more generally: impedance) we reach the important conclusion that charging effects are *strongly determined by the resistance or impedance of the environment* (i.e. leads and source). In particular these effects are most prominent if this impedance is large.

Before continuing with some more of the theory it is worthwhile to step back and consider the experimental implications following from the requirement of a high-resistance environment. It requires a resistor of resistance $R\sim100$ kΩ or more, with the additional requirement that the capacitance of the resistor (which is intricately connected to the geometric size of the resistor)) has to be considerably smaller than the capacitance of the junction, as it otherwise will increase to total capacitance of the system and thus reduce the charging effects (eq. (7.3)). From eqs. 7.1 we derive that this can only be realised if the size of the resistor is kept small, typically <1 μm. The combination of such a large-value resistor with a small size is by no means obvious, both from a technological point of view (see Appendix A) as well as based on more fundamental aspects (requiring the resistor to be made of a material of very large resistivity, while still being metallic (see the discussion following eq. (2.5))!

7.2.d. *Single electron charging in systems containing tunneljunctions and capacitors.*

For single-electron charging effects to occur we require that the charge is well localised at a small metallic conductor. This requires that the tunneljunction is connected to a high impedance "environment", or to inhibit the flow of electrons through the connecting element all together. In the first case one can employ a second junction as the "load" for the first one. Evidently this argument holds also vice-versa: the first junction acts as a high resistance for the second junction as well. In the second case a (common) capacitor can be taken as the connecting load. We will start discussing the second case, and show

that this logically leads to the first. In particular we concentrate on the resulting energy balances in the circuit.

Fig 7.6a shows the diagram for the circuit comprising the series connection of a tunneljunction (parameters R and C_r) and the capacitor of capacitance C_g. The reason for the two subscript names r and g will become clear once we transform our present circuit to the double junction variety. In addition a voltage source of potential V_g is connected to the capacitor C_g. We want to derive the energy of the total system (i.e. capacitors and voltage source). More specifically, we are interested in an expression for the energy that allows us to evaluate the energy differences that are associated with the tunnelling of electrons in the system, i.e. in this specific case from the zero-voltage (or ground) lead through the tunneljunction onto the island formed by it and the capacitor. Basically, this is the *free energy U* of the system *for the exchange of electrons with its environment.*

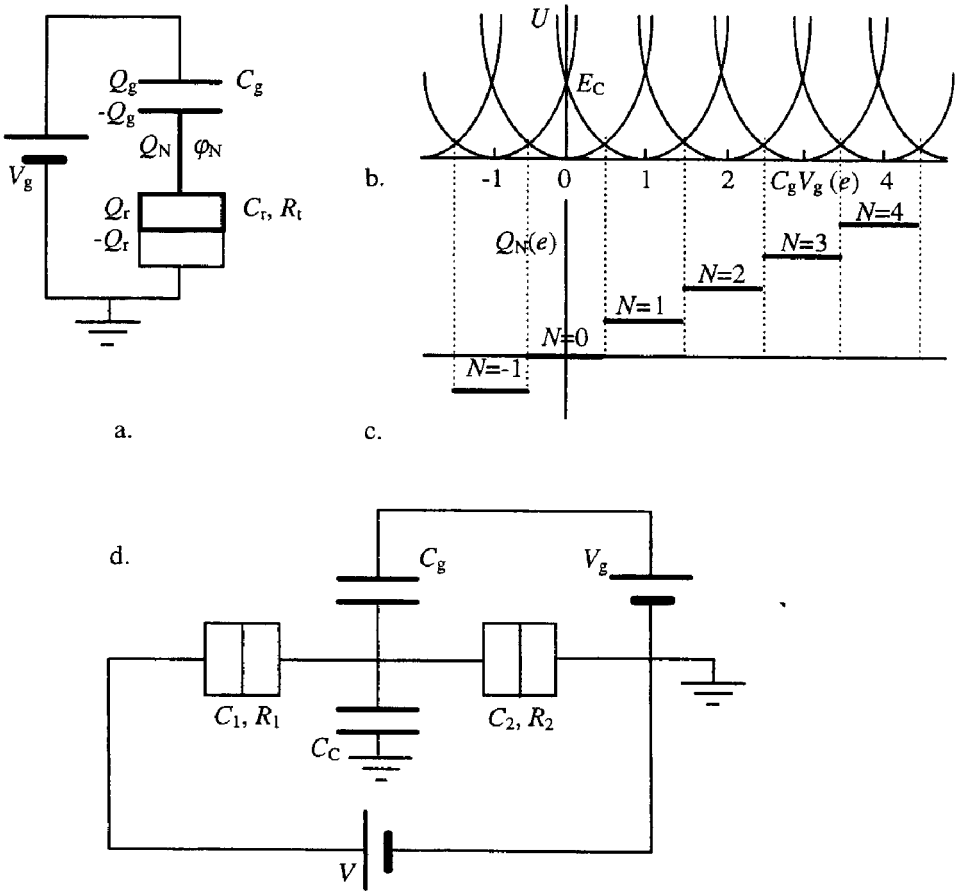


Figure 7.6. Series circuit comprising a tunneljunctions and a capacitor (a.); the charges on the two capacitive elements as well as the potential of the central island, φ_N, are indicated. b. The free energy of the system of a. in dependence of the applied voltage V_g, showing a periodic behaviour with respect to the addition (or removal) of one electron charge |e| *to the island charge Q_N, as shown in c. In d. the diagram of the double junction circuit is shown; for V=0 this circuit can be redrawn to circuit a.*

The energy is built up from three different contributions, two of these being electrostatic in nature and one being electrochemically, resulting from the work done by the voltage source. Evidently, as we are studying single-electron charging effects, all capacitors are small, making the charging energy $E_C=e^2/2C_\Sigma=e^2/2(C_g+C_r)$ the relevant energy scale following eq. (7.3). This implies that at the island formed by the two capacitors the total charge has to be due to an *integer number* of electrons, i.e. $Q_N=-Q_g+Q_r=Nel$.
Let us first assume that no net charge sits at the island, or $Q_N=0$ and $N=0$. This implies $Q_g=Q_r$. The application of the voltage V_g to the circuit effectively charges the two capacitors in series yielding the effective capacitance $C_s=C_gC_r/C_\Sigma$. Thus the first term of the energy equals

$$U_1 = \int_0^{V_g} VdQ = \int_0^{V_g} C_s V dV = \frac{C_s V_g^2}{2} = \frac{C_r C_g V_g^2}{2C_\Sigma} \qquad (7.13a)$$

Now we allow charge (say, N electrons) to tunnel from the island to zero potential through the junction C_t. The resulting charges on the two capacitors are denoted by Q_{Ng} and C_{Nr}, and the total net charge on the island becomes $Q_N=-Q_{Ng}+Q_{Nr}=Nel$. This added (or removed) island charge leads to the other two contributions to the energy.
The isolated charge $Q_N=Nel$ sitting at the total capacitance C_Σ leads to the straightforward electrostatic contribution

$$U_2 = \frac{Q_N^2}{2C_\Sigma} \qquad (7.13b)$$

i.e. the second energy term.
In addition, the removal of the electrons from the island not only affects the total charge of the island, but evidently also the charges present at the two constituting capacitors. The additional charge at the capacitor connected to the voltage source is denoted by Q_{Ng}. It can be straighforwardly shown to be $Q_{Ng}=(C_g/C_\Sigma)Q_N$, and so the work done by the source to transfer this charge from the + to the - terminal (=ground) is the third contribution to the energy,

$$U_3 = Q_{Ng}V_g = \frac{C_g V_g}{C_\Sigma} Q_N \qquad (7.13c)$$

As mentioned before we are interested in effects on the energy resulting from tunnelling events, i.e. if the charge on the island changes by a unit *el*. This implies that the $Q_N=0$ term of eq. (7.13c) can be suppressed. Thus summing and rearranging the terms of eqs. 7.13b & c yields

$$U' = \frac{(Q_N + C_g V_g)^2}{2C_\Sigma} - \frac{(C_g V_g)^2}{2C_\Sigma} \qquad (7.14)$$

While also the last term of this expression does not depend on Q_N we may suppress it, thus arriving at our final expression for the free energy

$$U = \frac{(Nel+C_g V_g)^2}{2C_\Sigma} \qquad (7.15)$$

Note that in this expression C_gV_g acts as a continuously variable charge added to the net integer charge $Q_N=Nel$ on the island, *without actually changing* Q_N: *it only affects the*

energy of the system via a polarisation charge and a consequential redistribution of the total charge Q_N over C_g and C_r.
Figure 7.6b schematically shows the parabolic dependence of eq. (7.15), with each parabola being associated with an integer "island filling number" N=..., -2, -1, 0, 1, 2, ... From this diagram it is clear that, by varying the control voltage V_g, each parabola can be traced, provided the charge at the island is kept fixed. It is, however, quite clear that for each value V_g the free energy (7.15) is at its minimum for only one unique island charge Q_N. So we anticipate that, if the system is allowed to stay at the lowest energy, i.e. the ground state, (see below) the number of electrons at the island will change one by one at increasing (or decreasing) control voltage, each time we cross the next lying parabola. Figure 7.6c shows this stepwise change of the island charge in dependence of the voltage, with each step having a size $\Delta Q=|e|$; a similar behaviour is shown in the top panel of figure 7.7.
Evidently also the potential of the island will be affected by the charge that is accumulated on it, as well as by the control voltage. It is simple to show that the electrostatic potential of the island containing N (excess) electrons and subject to the external potential from the gate via the coupling capacitor C_g, behaves like

$$\varphi(N,V_g)=\frac{N|e|+V_gC_g}{C_\Sigma} \qquad (7.16)$$

So, we see that the circuit of figure 7.6a allow us to control the number of electrons on an isolated conductor that is weakly connected to leads to the level of a *single electron* and this can be done in a very *precise and predictable* way, provided that we meet the conditions imposed by eqs. (7.3) and (7.6). Stressing this quite outstanding capability this circuit, comprising one tunneljunction in series with a capacitor, is called a *single electron box.*

It is very simple to extend this simple two-terminal circuit (basically consisting of the ground lead and the lead to which the control voltage is connected) to a three-terminal variety. Figure 7.6d shows the resulting diagram. It includes two series connected tunneljunctions of capacitances C_1 and C_2 , the island with a capacitance C_c to ground , and a capacitive coupling C_g that connects to the control voltage V_g.
The energy of the resulting circuit can be obtained in a straightforward manner from eq. (7.14), which effectively comprised a term due to the integer charge at the island added to the contribution related to the work performed by the (external) voltage source(s) while reaching the charged state, or

$$U(Q_N,V_g,V)=\frac{Q_N^2}{2C_\Sigma}+\sum_{sources}Q_sV_s=\frac{Q_N^2}{2C_\Sigma}+\frac{Q_N}{C_\Sigma}(C_gV_g+C_1V)+U_0 \qquad (7.17)$$

with Q_S being the charges transported through the sources V_s. While now the control voltages (connected via true capacitors) and the bias voltages (connected via tunneljunctions) are include on an equal footing, this expression allows us to study linear (V_{bias} very small) well as non-linear (V_{bias}~$E_C/|e|$) properties like I-V characteristics. Concentrating on the double junction circuit of figure 7.6d we will discuss these two cases now, and see how the current that flows through the two junctions and the bias source is affected by the bias source voltage V and the control or gate voltage V_g.
A. The small-bias linear regime ($|V|<<E_C/|e|$)

From eq. (7.10) we derive that, at zero temperature, tunnelling of particles onto or out of the island will be possible whenever this will lead to a decrease of the total energy U, i.e.

$$\Delta U \equiv U(Q_N \pm |e|, V_g) - U(Q_N, V_g) \leq 0 \qquad (7.18)$$

Using eq. (7.15) this condition is equivalent to $|C_g V_g + N|e|| > e/2$. Consequently the system is in a stable charge state only if

$$-(N+1/2)|e| < C_g V_g < -(N+1/2)|e| \qquad (7.19)$$

with $Q_N = -N|e|$

Equation (7.19) states that, at a given gate voltage, the system will adjust the net charge at the island such that this stable state is established. In figure 7.6b these condition points are indicated by the dashed vertical lines, i.e. they occur at the crossing of two neighbouring parabolas that differ by one unit charge. Note the periodicity of the control

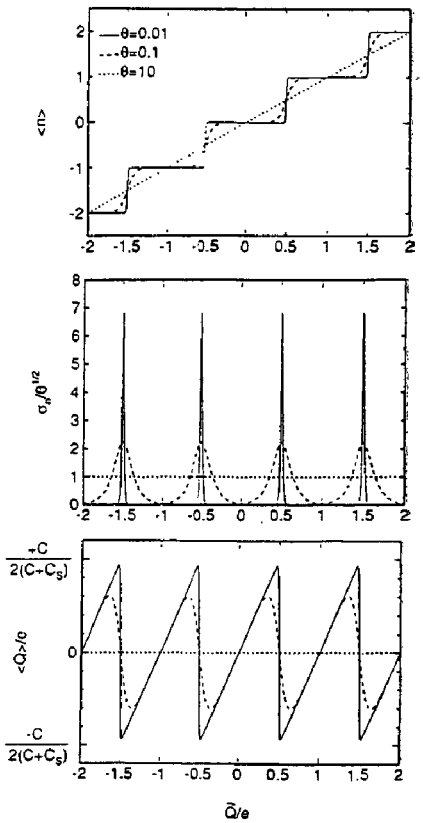


Figure 7.7. The charging state of a double junction circuit, in . dependence of the gatevoltage V_g. a. The (average) charge Q_N, showing a clear stepwise behaviour. b. The conductance, which is strongly suppressed if the charge Q_N is fixed, i.e. at the plateaux of a. c. The potential $\phi(N)$ of the central island, with a stepwise decrease if the charge Q_N changes and a linear rise with V_g at constant charge. The solid line is at a temperatures much smaller than E_C/k, while the dashed and dotted lines are obtained for increasingly larger temperatures.

voltage $\Delta V_g = e/C_g$. In case the system is in an unstable state (either because too many or too few electrons reside at the island to meet the condition (7.19)), a number of tunnel events will occur to reach the stable state, after which no further exchange of particles with the reservoirs will occur.

In general the system now contains a fixed number of electrons at the island, and no further transport is allowed, i.e. I=0: the double junctions system is in the *Coulomb blockade* state. However, if V_g matches the condition $U(Q_N \pm e, V_g) = U(Q_N, V_g)$, the degeneracy of the energies of the two states (i.e. ΔU=0) differing by one unit charge $|e|$ implies that the system does not have any "preference" for containing N or $N \pm 1$ electrons. Thus electrons are allowed to enter and leave the island, i.e. electron transport may take place. E.g. for V_g=-(N-1/2)$|e|/C_g$ at a small bias voltage the system will evolve in time via the sequence of charge states Q_N=-$N|e|$, -(N+1)$|e|$, Q_N=-$N|e|$, -(N+1)$|e|$, ... The sequential addition and leave of single electrons to the island leads to a non-zero current flowing due to the (small) bias voltage V, i.e. the conductance G=I/V is non-zero at the points. This is seen in the middle panel of figure 7.7. The sequence of peaks and valleys in dependence of the control voltage V_g is denoted as *Coulomb blockade oscillations*. The behaviour of the potential of the island, following from eq. (7.16), is shown in the lowest panel of the same figure. Note that (at zero temperature) only *one electron at a time* can be added to the island: the repulsive Coulomb interaction prevents more particles to sit on the island, in this way again forcing the electronic transport to become *correlated in time* (compare to figure 7.5).

Problem: discuss the time dependence of the potential of the island, ϕ, when the conductance is at its maximum and a current I is transported.

For the present case of small bias the degeneracy condition $U(Q_N \pm e, V_g) = U(Q_N, V_g)$ for tunnelling can be interpreted also in a different and more fundamental way. From thermodynamics we know that in a system where particle exchange is allowed, *thermodynamic equilibrium will be reached* whenever the *electrochemical potential throughout the system is constant*. So, in the case that the tunnelling degeneracy condition ΔU=0 holds transport is allowed and particles are exchanged throughout the system, including towards the particle reservoirs (or leads) connected to the island. With the leads at electrochemical potentials μ_1 and μ_2 respectively (with μ_1-μ_2=-$|e|V$ is small and μ_1, $\mu_2 \cong E_F$), this implies the following equivalent

$$U(Q_N \pm e, V_g) = U(Q_N, V_g) \Rightarrow \mu_1, \mu_2 = \mu(Q_N, V_g) \qquad (7.20)$$

This is a very important and general result. Rephrased in words is states that tunnelling in a charging system occurs whenever the *electrochemical potentials of all the states involved are aligned.*
Note that this condition for alignment of the electrochemical potentials implies that *Coulomb blockade is suppressed*, as particle exchange is the main consequence of the alignment.

We now want to discuss a few experimental results representative for the linear regime at low temperature, i.e. $|eV|, kT << e^2/C_\Sigma$. The first one concerns a quantum dot in a 2DEG, for which Fig 7.8 shows a typical example. By applying negative gate voltages the electrons are confined to the 7 "white" areas delimited by the gates A-E. The two tunnel barriers are formed by the two gate pairs A,F and C,D, while the island is formed in the centre by these gates as well as the two middle gates indicated by B and E. Note that the voltages on the gates are chosen such as to close the long, narrow channels between the gates A,B, C,B as well as between F,E and D,E by depleting the 2DEG in between. The conductance of the barriers can be adjusted independently by changing either $V_{A,F}$ or

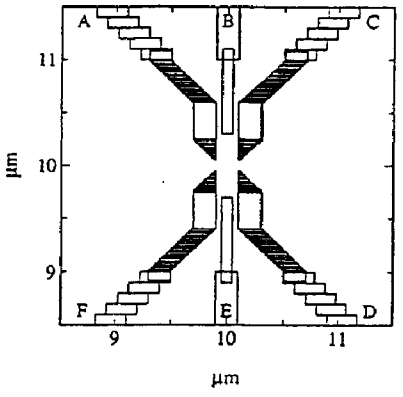


Figure 7.8. Gate pattern on the surface of a GaAs heterostructure Negative voltages on the gates suppresses the electron density underneath the gates, leaving a finite value in the areas delimited by the dashed lines. Gate pairs A,F and C,D induce the tunnel barriers, while gates B,E close the dot and control the energy and the total charge Q_N.

$V_{C,D}$. The gates B,E are used to control the electrostatic energy of the electrons residing on the island (or dot), thus acting as the control capacitance C_g. Dots of this type have typical sizes of a few 100 nm contain a few hundreds of electrons and have a capacitance C~100 aF (1aF=10^{-18}F), leading to a charging energy E_C~1meV~10K. Figure 7.9 shows the current through a similar device, in dependence of the voltage V_{GE} on gate E. With the applied bias voltage V=50 μV the vertical scale represents the conductance through the device. It nicely displays the *Coulomb blockade oscillations* in accordance with the theoretical curves shown in Fig. 7.7b. At this stage we will not discuss the mechanism leading to the strong "wavy" modulation of the size of the successive peaks.

Problem: discuss possible mechanisms leading to the general trend in figure 7.9 of the increase in conductance with increasing (i.e. less negative) gate voltage.

The second example of an experiment concerns a single electron box (see just below eq. (7.16) in a metallic junction system. Fig 7.10 shows the diagram of the system, basically comprising two coupled single-electron charging circuits. The first part, which is the actual system of interest, act as the electron box: it contains two metallic junctions and two capacitors C_S and C_C coupled to the central island b. Note that, while both junctions are connected to the zero potential, they effectively act as a single junction, of the type shown in figure 7.6a. The aim of the experiment is to measure the potential $\phi(N,U)$ of the

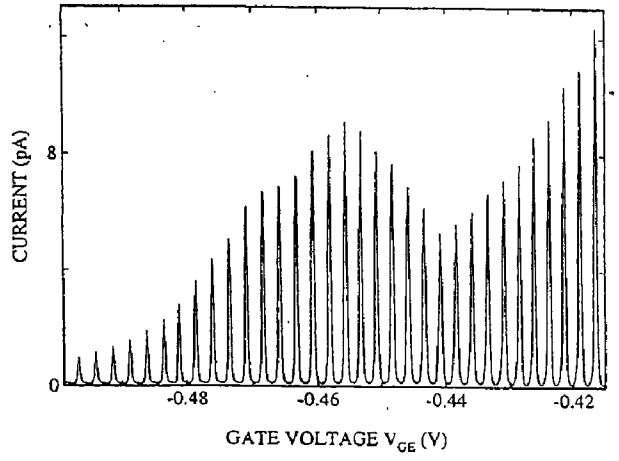


Figure 7.9. Coulomb blockade oscillations in the current through a quantum dot of figure 7.9 at a constant bias voltage V=50 μV, in dependence of the voltage the middle gate E, at a temperature T~20 mK

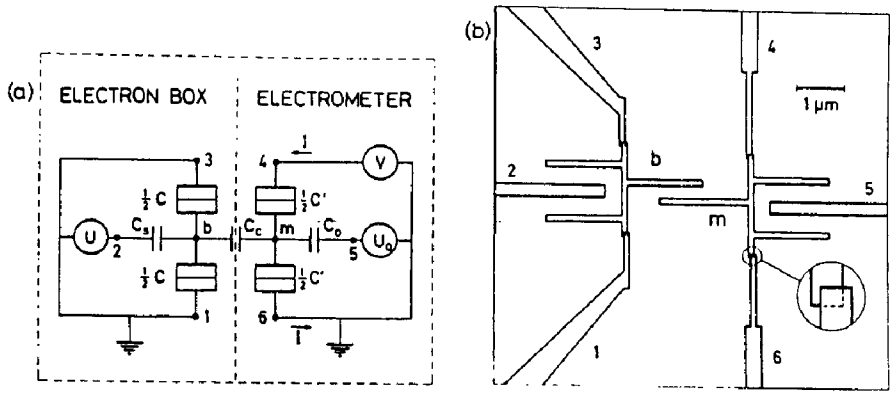


Figure 7.10. A circuit containing two capacitively coupled single-electron tunnelling devices, realised by using Al-AlOx tunneljunctions. It comprises one SET device (m, C') used as an electrometer to measure the potential and its associated charge located at the island of the second SET structure (b, C), called electron box. The coupling of the potential of the electron box island b is realised by capacitor C_C. a. shows the basic circuit diagram; b. shows a drawing of the actual structure.

central island *b* in dependence of the voltage *U* on gate 2 (figure 7.10a). To this purpose the second double junction with island *m* acts as an (extremely sensitive) charge detector or electrometer. It consists of the junctions indicated by *C'*/2 and the gate C_O. The precise operation of this electrometer (in the non-linear regime) will be discussed below. C_C acts as the coupling between the two central islands *b* and *m*. In this way the potential $\phi(N,U)$ of the central island *b* affects the potential of island *m*, which in turn controls the Coulomb blockade of this second circuit, i.e. it determines the current *I* through the electrometer single-electron device. Some characteristic values of the Al/AlOx-based structure are: the junctions *C* and *C'* have an AlOx barrier with a size of ~50*50 nm^2; $C=C'\cong 600$ aF and $C_S=C_O\cong C_C\cong 70$ aF, and the junction resistances are ~600 kOhm, i.e. $>>h/2e^2$ (=25 kOhm). Fig 7.11 shows the experimental result obtained at a temperature

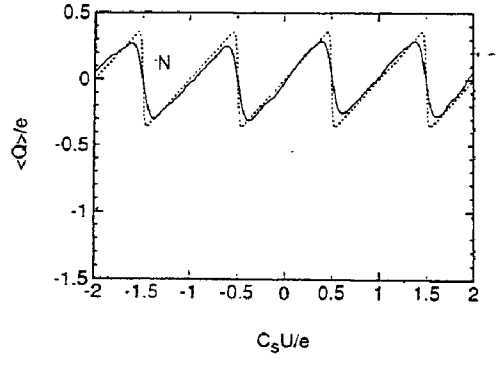


Figure 7.11. The electrostatic potential $\phi(N,U)$ induced by the gate voltage U. The gradual increase results from the continuous polarisation charge at a constant number of electrons; the steep decrease indicates the stepwise change of the charge by one electron.

of ~20 mK. Comparing this curve with the theoretical one (Fig 7.7c) demonstrates good agreement, nicely showing the sawtooth shape. The linear ramp is associated with the condition of a fixed number of charges and the steep drop demonstrates the switching of this net charge with one unit. The finite slope of the steep drop is a consequence of the finite temperature.

B. The finite bias non-linear regime.

Now we want to continue with the theory and concentrate on charging effects in the double junction circuit of figure 7.6d in the non-linear regime, i.e. $kT<<e^2/2C_\Sigma<|eV|$. Using eq. (7.17) one immediately can calculate the free energy differences associated with the tunnelling of electrons onto or from the island. For reasons of simplicity we modify the circuit of fig 7.6d such as to make it more symmetric, by applying the bias voltage V in a symmetric manner, i.e. the left and right lead are positioned at potentials $V_1=V/2$ and $V_2=-V/2$. The tunnelling of a unit of charge through junction 1 yields

$$\Delta U(1,N,N+1) \equiv U_{final} - U_{initial} = \frac{e}{C_\Sigma}((N+1/2)e + C_g V_g - V_b(C_2 + C_g/2)) \quad (7.21)$$

This event increases the island charge from N to N+1. In the same way we can obtain three more energy differences for tunnelling of one unit charge $|e|$ out via junction 1 (N =>N-1), the addition of $|e|$ via junction 2 (N=>N+1), and the loss of it via the same junction (N=> N-1). Equating the four differences to zero defines the conditions for the various tunnelling events.
These conditions can be represented graphically by four lines in the (V,V_g)-plane. Figure 7.12 shows this so called *stability diagram* of the double junction device. Inside the array of diamonds along the V_g axis symmetrically around V=0 the net charge Q_N is stable, indicated by the integers ...,-1,0,1,2,..., and the system is blocked. In accordance with intuition, the range of gate voltage V_g over which the blockade is maintained decreases with an increase of the bias $|V|$. Outside the diamonds the Coulomb blockade is overcome by the bias voltage V and tunnelling can take place. To investigate this process in more detail one can quantitatively evaluate the tunnel currents, using the expressions derived for the tunnelling rate through a single junction (eqs. (7.9) and (7.10)). For the double junction case four Γ's have to be evaluated using the energy differences $\Delta U(k,N,N')$ given in eq. (7.21) and in the text immediately following it. From these rates one immediately

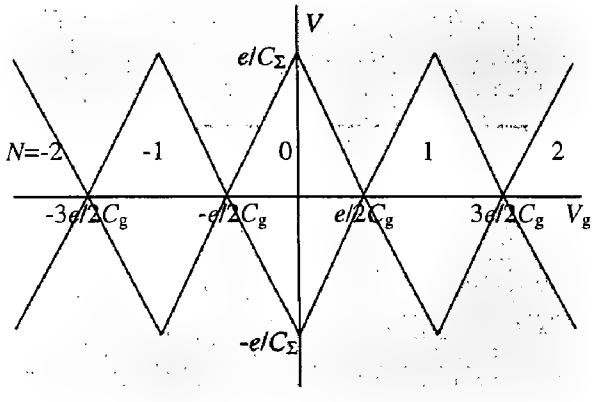


Figure 7.12. Stability diagram for the double junction circuit of fig. 7.6d. $C_1=C_2$, $C_\Sigma=C_1+C_2+C_g$ and $C_c=0$.

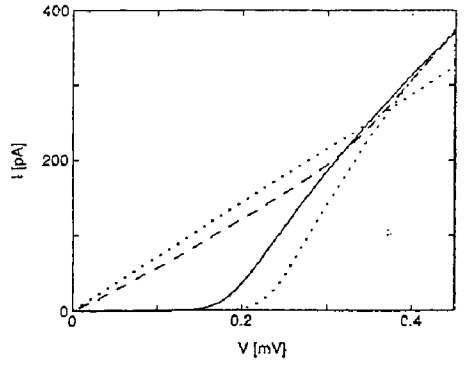
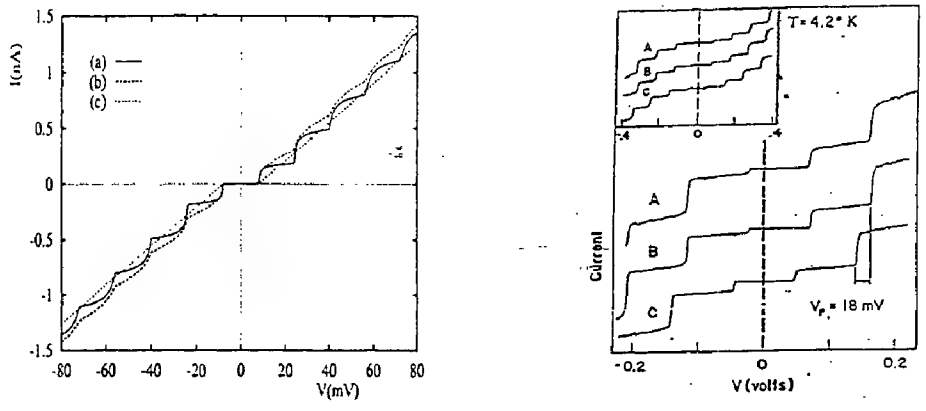


Figure 7.13. I-V curve of a metallic double junction device. The characteristics of the junctions are R=350 kΩ and C=0.3 fF. The solid and dashed curves are experimental results at maximum and minimum Coulomb blockade respectively.

obtains the overall tunnel rate Γ(1=>island=>2) from electrode 1 via the central island into electrode 2.
Fig. 7.13 shows an experimental I-V-curve for a metallic double junction circuit, obtained at 20mK for gate voltages V_g=0 (maximum Coulomb blockade) and V_g=$e/2C_g$ (minimum CB), compared to a numerical evaluation using eq. (7.21). Note the quite good agreement between the simulation and the experiment. Deviations, in particular at higher current levels, are likely to be due to the increase of the effective electron temperature at the central island.

In the preceding discussion we have assumed that the two junctions are identical in resistance R_t and capacitance C_j, i.e. the *RC*-leakage time τ=R_tC_j is the same for both junctions. If these τ's differ strongly the evaluation of the rate equations shows that the I-V characteristic develops a sequence of steps, which is called a *Coulomb staircase*.



a. b.

Figure 7.14. The Coulomb staircase. a shows the calculated I-V curves for the ratios {R_1/R_2;C_1/C_2} equal to {25,10}(solid), {25,1}(dashed) and {25,0.1}(dotted). b. shows an experimental result obtained by STM on an In droplet at 4.2K.

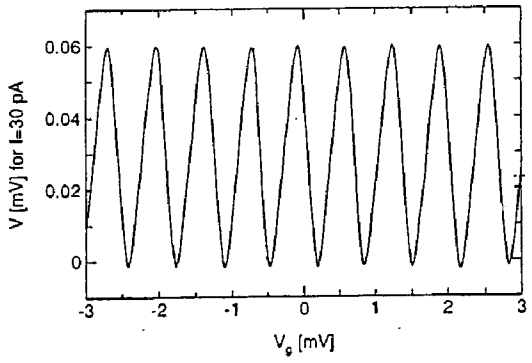


Figure 7.15. Characteristics of a current-biased single electron electrometer, at T~20 mK. Biased at 30 pA the voltage gain is ~0.35.

Figure 7.14a shows this behaviour for the case of three ratios of the R and C parameters $\{R_1/R_2;C_1/C_2\}$. It clearly shows that the staircase becomes more prominent with increasingly different time constants of the two junctions. Figure 7.14b shows an experimental result obtained from a scanning tunnelling STM experiment on a small In droplet at 4.2K, showing a magnificent Coulomb staircase. Employing the same formalism based on the rate equations (7.9) and (7.10) and the energy differences in the double junction circuit we also can understand the operation of the electrometer used in the electron box experiment of figure 7.10. This electrometer (i.e., the circuit at the right) simply is a double junction device operating in the non-linear regime. Looking again to figure 7.13 it is obvious that, if the device is biased at a constant current, the resulting voltage across the device will be modulated by the potential of the electron box island that couples to the electrometer device via capacitor C_C. E.g. biased at I=50 pA it will lead to approximately 0.22 mV at maximum Coulomb blockade and 0.10 mV at the minimum. Fig. 7.15 shows this modulation for a similar device operating at 30 pA current bias. Note from the very low noise level on the V signal that this charge detector has a sensitivity *far better than a unit charge*. More detailed measurements show values approaching 0.0001|e| (for 1 sec. detection time), making the SET electrometer at least three orders of magnitude more sensitive than any other existing charge detector!

7.2.e. *Single-electron charging in islands with a discrete energy spectrum, or Quantum Dots*

In the preceding sections charging effects have been discussed under the constraint that the electrons behave largely classically, i.e. neglecting their wave nature. This is fully adequate if the central island contains a large number of free electrons (note: this concerns the *total number of electrons*, which needs to be clearly distinguished from the *number of excess electrons*, associated with Q_N!!). In this case the confinement of the electrons in the island will not lead to well-determined zero-dimensional (0D) single electron states ("particle in a box", see subsection 1.2b), as the typical 0D energy splitting between levels will be much smaller than the width of each level (determined by the phase coherence time and the leakage time through the barriers out of the island into the connecting reservoirs). The strong overlap of the levels leads to an effectively continuous density-of-states (DOS). This is the typical case for metallic systems, containing a million or more electrons.

In the case of a semiconductor island the reduced electron density and associated large Fermi wavelength may lead to the opposite regime where the confinement results in well-determined 0D eigenstates for the electrons, with eigenenergies E_p. The fact that the properties of the island are determined by the quantum nature of the electrons has lead to the common use of denoting these islands as *Quantum Dots* (QD's).
The discrete DOS has a number of effects on the charging phenomena discussed before. In particular it leads to deviations from the e/C_g periodicity in V_g of the Coulomb blockade oscillations, and the phenomenon of resonant tunnelling through the device. In addition in the non-linear regime of large bias voltage the Coulomb oscillations become multiply split, with the structure in the split peaks providing a method to obtain a good quantitative estimate of the 0D level splitting. These aspects will be discussed in the next paragraphs.

A. The small bias linear regime ($|eV| \leq E_C, E_{0D}$)

We denote the 0D energy levels of the quantum dot by E_p, enumerated by p=1,2,..., starting from the lowest level. The free energy for the metallic case, given by eq. (7.15), has been derived under the assumption of a large, continuous DOS throughout the whole structure (island and leads). Assuming that the leads have a continuous DOS, the total free energy of the system with a *discrete* DOS in the island is obtained as the sum of the energies of all *occupied* 0D states added to the free energy of eq. (7.15) or (7.17). For the small bias case (eq. (7.15)) this yields

$$F(N,V_g) = U(N,V_g) + \sum E_{0D} = \frac{(Q_N + C_g V_g)^2}{2C_\Sigma} + \sum_p^M E_p \qquad (7.22)$$

with the summation running up to the highest occupied state E_M. Note again that N denotes the *excess* charge in the dot, while M counts the *total* number of electrons inside the dot, the majority of which is electrostatically compensated by the positively charged ionised background donors and/or an external gate potential (see chapter 2).
In the case of removal (or addition) of an electron from the island the change in energy now not only originates from the electrostatic contribution (7.15), but in addition the *difference in 0D energy* ΔE_{0D} has to be taken into account. Assuming the N-th electron to occupy state p, (and similarly state p-1 is occupied by the N-1-th electron) the energy degeneracy condition (eq. (7.18)) $\Delta U(N,N\text{-}1)$=0 at a given V_g to be replaced by

$$\Delta F(V_g,N,N\text{-}1)=F(N)-F(N\text{-}1)=\Delta U(N,N\text{-}1)+(E_N-E_{N\text{-}1})=\Delta U(N,N\text{-}1)+\Delta E_{0D}=0 \qquad (7.23)$$

Here we have rewritten E_p=>E_N, etc. Note that for the case of a metal $E_N \cong E_{N\text{-}1} (\cong E_F)$, and so (7.23) reduces to the former result ΔU=0 of eq. (7.18). The degeneracy condition (7.23) leads to suppression of the Coulomb barrier, allowing particles to exchange freely throughout the system, thus leading to a maximum in the conductance through the dot.
Following the introduction of eqs. (7.18) and (7.19) we have seen that for the classical metallic system the (excess) charge Q_N is stable for a gate voltage interval given by $\Delta V_g = e/C_g$, i.e. it is independent of the number N of electrons residing on the island. Thus, the peaks in the Coulomb blockade oscillations in the conductance show a regular pattern spaced at this fixed voltage interval (see figure 7.7b). This no longer holds for the case of the quantum dot. In general in a quantum dot the spacing between the various energy levels will depend on the specific states we are dealing with, i.e. $(E_{N+1}-E_N) \neq (E_N-E_{N\text{-}1})$ with the first difference associated with N+1 and N electrons and the second with N and N-1. This implies that the required change in gate voltage to go from the degeneracy

condition $\{N+1,N\}$ to $\{N,N-1\}$ differs from that in going from $\{N,N-1\}$ to $\{N-1,N-2\}$. Consequentially the peak spacing in the Coulomb blockade oscillation pattern will not be constant. Figure 7.17a shows this behaviour.

As discussed at the end of section 7.2.d we know that the thermodynamical consequence of the free exchange of electrons dictates that the electrochemical potentials of the quantum dot and the reservoirs have to align, thus yielding

$$\mu(N,V_g)=\mu_1=\mu_2 \qquad (7.24)$$

Using this notion of alignment allows a significant simplification of pictorially presenting the transport of electrons through a single quantum dot. We will return to this point when discussing figure 7.16.
The electrostatic potential $\varphi(N,V_g)$ of the dot will be given again by eq. (7.16). It is clear that the occurrence of 0D energy levels in the dot determines the chemical (or kinetic) energy of the system. As the electrochemical potential is defined as the sum of the electrostatic potential and the kinetic (or chemical) energy, 0D levels will affect the electrochemical potential $\mu(N,V_g)$, leaving the potential $\varphi(N,V_g)$ unaffected. More precisely, the addition of a single electron to the dot at fixed gate voltage yields changes in φ and μ given by

$$\Delta\varphi=\varphi(N+1)-\varphi(N)=e^2/C_\Sigma \qquad (7.25a)$$

$$\Delta\mu=\mu(N+1)-\mu(N)=E_{N+1}-E_N+e^2/C_\Sigma=\Delta E_{0D}+e^2/C_\Sigma \qquad (7.25b)$$

From eq. (7.25b) it is obvious that the change in μ due to the change of the number of electrons on the dot is always larger than the change of the electrostatic potential φ. In contrast in a metal the very large DOS makes $E_{N+1}=E_N$ and so in this case we do not need to discriminate between the behaviour of φ and μ.
Fig. 7.16 shows the potential landscape of a dot, assuming the difference of electrochemical potentials of the leads $\Delta\mu=\mu_1-\mu_2(=eV_{bias})$ is very small (compared to E_C

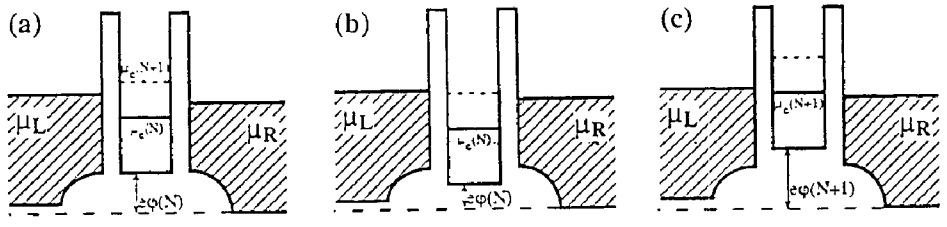


Figure 7.16. Schematic energy landscape representing a quantum dot. The electrochemical and the electrostatic potential are indicated for the cases of Coulomb blockade (a) and transport condition without (b) and with (c) the N+1th electron inside the dot.

and ΔE_{0D}). The position of the electrochemical potential of the dot, $\mu(N)$, is shown in dependence of the (not shown) gate voltage, with the potential increasing with more negative gate voltage. In a. the system is blocked as $\mu(N)<\mu_{1,2}$ and $\mu(N+1)>\mu_{1,2}$. By making V_g less negative, $\mu(N+1)$ can be aligned with $\mu_{1,2}$ enforcing the energy degeneracy condition for N and $N+1$ electrons, and so sequential single electron transport can take place leading to a finite conduction, as shown in b. (with N electrons in the dot) and c. (containing $N+1$ electrons).

Fig 7.17 shows the behaviour of the conductance G, the electrochemical potential $\mu(N)$ and the electrostatic potential energy $e\varphi(N)$ of the Coulomb island, in dependence of the gate voltage V_g. In 7.17a the by now familiar Coulomb blockade oscillations are displayed. At zero temperature and infinitely weak coupling to the leads the width of the peaks approaches zero (provided no intrinsic life-time limiting mechanisms exist for the electrons in the QD!). From the behaviour of μ and φ (7.17b and c.) two new aspects become apparent. From the requirement of the alignment of $\mu(N)$ and $\mu_{1,2}$ for the peak conductance (compare a. and b.) and the different values of consequetive 0D energy level splittings $\Delta E_{N,N-1}$ we immediately conclude that the intervals ΔV_g between the peaks in the conductance also will become unequal, in contrast to the metallic case (compare to Fig. 7.7!). So, the variation in gate voltage intervals *directly reflects the fluctuations in the 0D level splitting*. Stated differently, from the variations in gate voltage intervals we are able to directly obtain the *energy spectrum* of the electrons contained in the QD. This also leads to the second point. From eq. (7.25b) we see that the electrostatic potential $\varphi(N)$ jumps by e^2/C at the addition of one electron (i.e. at the point of degeneracy), leading to the saw-tooth shape of fig 7.17c. However, it also has to compensate for the *additional* contribution of the 0D splitting. This leads to the *overall decrease* of the potential φ with gate voltage.

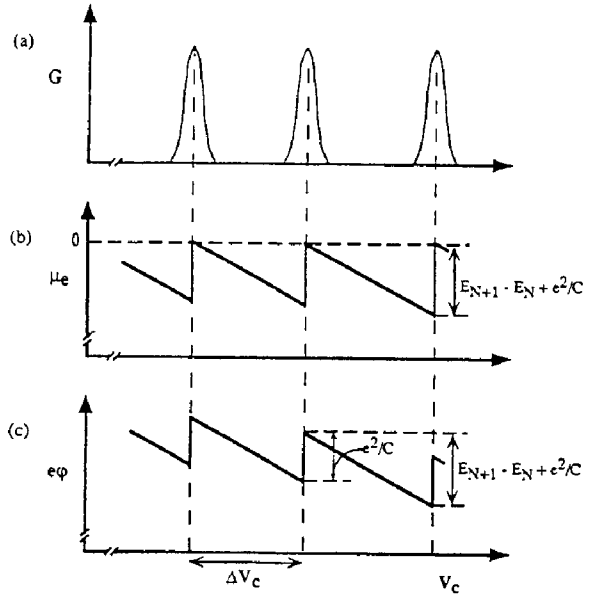
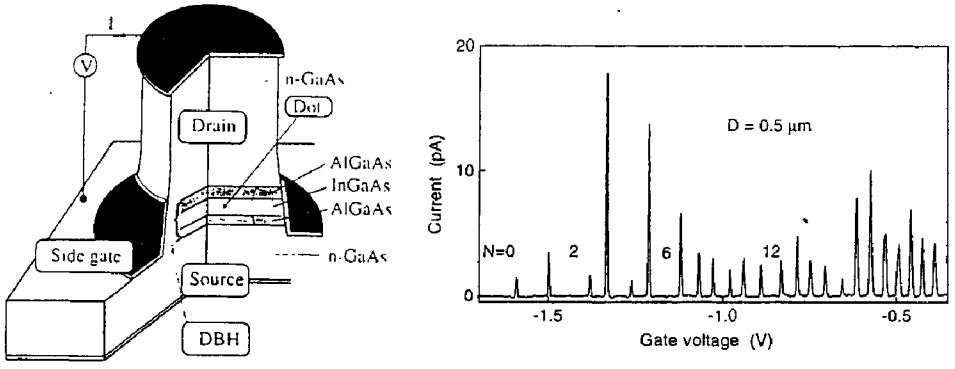


Figure 7.17. Coulomb blockade oscillations in a quantum dot. a. Oscillations of the conductance versus gate voltage. b. Saw-tooth oscillations of the electrochemical potential c. Saw-tooth-like oscillations of the electrostatic potential showing the effect of discrete 0D dot states.

Before turning to some experimental results, we want to regard the various energy scales in the problem from a slightly different perspective. To do so, we may look to a quantum dot as a system with a well-determined number of electrons that is quantum mechanically bound to it. Based on similarity we may redefine such a system as an *artificial atom*, even though the precise way of binding the electrons is not the same: via an attractive Coulombic force from a charged nucleus in a real atom, versus a repulsive Coulombic force from a gate in a QD. Despite this difference, which will affect the details of the binding potential, the similarity is striking. It is for this reason that also the *addition energy* for adding (or removing) of a single electron to (or from) the QD can be seen as the equivalent of the *ionisation energy* of an atom. Just as well, the level splitting ΔE_{0D} is equivalence to the *excitation energy*. However, the most striking feature for these artificial atoms is that the variation of the gate voltage provides one with the unique opportunity to *in situ control the atomic number of the atom*!

Fig. 7.18 shows an attractive experimental result demonstrating the effect of 0D states on the properties of Coulomb blockade oscillations. It is obtained on a small quantum dot in a 2DEG in a GaAs-AlGaAs-InGaAs-AlGaAs-GaAs heterostructure (S. Tarucha et.al, Phys Rev. Lett. 77, 3613 (1996)). In contrast to the type of semiconductor structures discussed up to now this QD is of a *vertical* nature. Shown in 7.18a, it is obtained by wet and dry etching, with the collar-shaped gate electrode formed via metal evaporation (for more details, see Appendix A). This particular configuration allows one to reduce the



a. b.

Figure 7.18. Single-electron Coulomb oscillations in a small quantum dot. In a. the structure of the dot is presented, showing the source and drain leads, and the actual dot formed approximately midways inside the collar-shaped side gate. In b. the CB oscillations in dependence of the (collar) gate voltage are shown. As indicated, for voltages V_g<-1.7V no electrons are contained in the QD, with the first electron entering at V_g=-1.7V. Note the strong variation in the gate voltage intervals between the sequential conductance (or current) peaks, imaging the variation in 0D level splittings.

number of electrons contained in the QD down to N=0.
Clearly the peaks shown in figure 7.18b are not equally spaced in V_g. This demonstrates that, in addition to a "simple" charging contribution, varying 0D level splittings occur in the electronic spectrum. At first sight it may seem that the intervals ΔV_g vary stochastically. However, a closer look shows that the intervals behave in a more regular way. In particular, the intervals supporting N=2, 6 and 12 electrons are evidently larger than average, showing that for these occupation numbers, a larger level splitting occurs, thus demonstrating an enhanced stability for these conditions. This is what is well known in atomic physics as the occurrence of a *shell structure*, where a closed shell configuration for the electrons shows an increased stability. The result of figure 7.18b is the first evidence that a similar effect occurs in a quantum dot artificial atom as well. To generalise it even further, it should be noted that similar stability considerations hold for atomic nuclei as well. We will return to these aspects in chapter 9 on clusters.

B. The finite bias non-linear regime

Figure 7.18b clearly demonstrates the effect of 0D confinement on the transport through the QD. In the final part of this section we want to investigate these 0D states in more detail, employing the bias voltage to perform some type of *spectroscopy*.
In the preceding part we have assumed the bias voltage $V_b=(\mu_1-\mu_2)/e$ to be small compared to E_C/e or $\Delta E_{0D}/e$. The natural question to pose is how the conductance $G=dI/dV_b$ will develop if the bias exceeds any of these energy scales. For the case of (the largest of the two,) E_C we have seen (see discussion related to Fig. 7.14) that increasing the bias voltage may lead to the Coulomb staircase. Taking the derivative of figure 7.14a shows that peaks in dI/dV_b will arise at the positions of the current steps, at bias voltages $\sim E_C/e$. The step in current reflects that an *additional path (or channel)* becomes available for the electrons to tunnel through the system, i.e. 2, 3, 4,... (note: still an integer number!) electrons at a time are allowed to traverse the Coulomb island. In addition, this step in the current (and peak in G) occurs whenever the *bias voltage exceeds the relevant energy*, in this case the charging energy. In this respect the bias voltage acts as a measure for the energy involved, i.e. it allows *spectroscopy* to be done.
Let us now discuss the effect of the 0D splitting. Figure 7.19 shows the potential landscape of a quantum dot. The narrowly spaced lines (two sets, drawn as solid and dotted respectively) represent the 0D levels, with the mutual spacing (solid-to-solid or

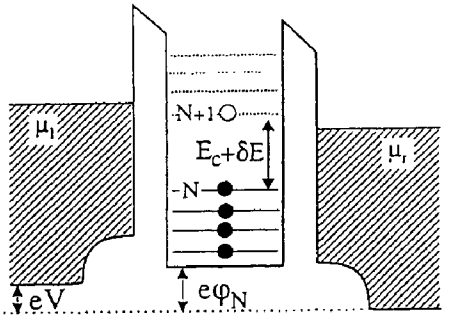


Figure 7.19. Potential landscape of a quantum dot, showing the the charging energy E_C and filled (closed dots) and empty 0D states of the dot.

dotted-to-dotted) representing the 0D splittings. In addition, the two sets are displaced to one another by the charging energy E_C. Note that in our (highly simplified!) picture each individual state given by $\mu(N)$ is built up from the straightforward summation of the electrostatic energy with the 0D confinement energy and so each charging state "carries" the *same* "ladder" of 0D states with it! As discussed before, the $\mu(N)$'s form the *addition spectrum* (i.e. resulting from the successive addition of single electrons) while the 0D ladder forms the *excitation spectrum* of the dot (i.e. keeping the number of electrons, and so the Coulomb contribution, constant). Evidently this is a considerable simplification, as a change of the occupation of 0D states (e.g., the electron being excited from E_N to E_{N+1}) will lead to a redistribution of the charge over the dot because of the different eigenfunctions associated with two distinct 0D states, yielding a change in the Coulomb contribution. This, selfconsistently, will modify the energy spectrum as such.

Figure 7.20 explains how electron transport through the dot takes place under the condition of non-negligible bias voltage. Similar to figure 7.19 we assume the 0D splitting to be smaller than the charging energy E_C. The bias voltage $V_b=-(\mu_l-\mu_r)/e$ is taken as $\Delta E < eV_b < 2\Delta E$. In the top part of figure 7.20 the gradually changing gate voltage pushes up the electrochemical potential of the dot via $\varphi(N,V_g)$. In the leftmost case (a) the system is (Coulomb) blocked: the state $\mu(N)$ lies below the lowest of the two reservoir electrochemical potentials, μ_r (as well as $\mu(N+1)>>\mu_l$). Note that in the ground state *all 0D states* E_N are occupied, as represented by the filled dots in the figure. In particular the 0D state E_{N+1}, i.e. the first excited state beyond those occupied in the ground state, will be empty, even though it seems energetically accessable as it is situated

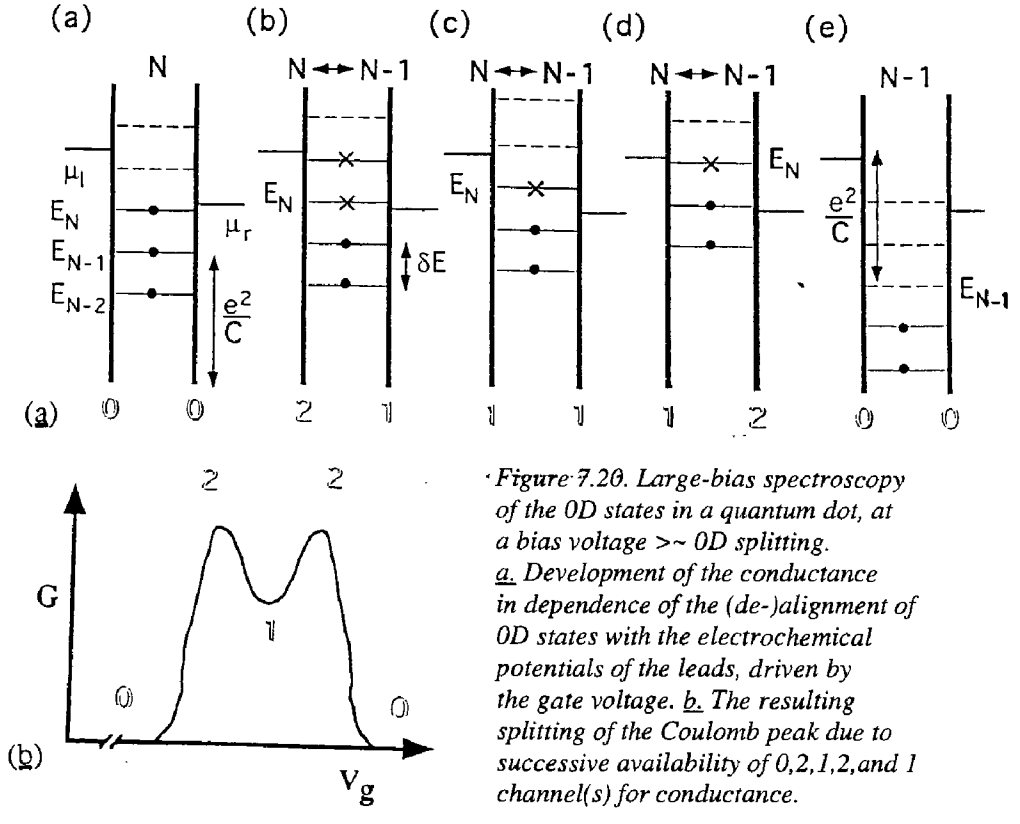


Figure 7.20. Large-bias spectroscopy of the 0D states in a quantum dot, at a bias voltage >~ 0D splitting. a. Development of the conductance in dependence of the (de-)alignment of 0D states with the electrochemical potentials of the leads, driven by the gate voltage. b. The resulting splitting of the Coulomb peak due to successive availability of 0,2,1,2,and 1 channel(s) for conductance.

between μ_l and μ_r: remember however that adding an electron will push the whole ladder upward by the charging energy, shifting the *total* energy of the state when occupied to well beyond μ_l.
Making V_g more negative (b) shifts the whole ladder upward, which will make $\mu(N)>\mu_r$ and so conduction will start. Note that in this case *two* parallel channels are available for the electron to traverse the dot: either the *groundstate* level $\mu(N)$ associated with the 0D state E_N; or the *first excited state* E_{N+1} which is just one 0D levelspacing higher in energy. It is crucial to note that *never the two states can be occupied simultaneously* due to the Coulomb interaction, and so still electrons are tunnelling sequentially in time!
It is interesting to note that the electron entering the excited state may chose from two possible ways to exit: either directly from this excited state or after relaxing to the groundstate. In the first case the tunnelling is said to be *resonant* and the phase of the electron will be preserved during the whole process, making it a single coherent quantum event. In the second case the electron loses energy while residing in the dot and the tunnelling is *sequential*: the phase will be lost and the entrance and exit tunnel process will not be coupled. Whatever the detailed mechanism, case (b) will lead to a strong increase in the current.
Continuing to case (c) the 0D spectral ladder is shifted upward so far as to bring the first excited state beyond μ_l, which will close this channel for transport and so only the groundstate channel is left to contribute. This will lead to a reduction of the total current carrying capacity of the system, i.e. a drop.
The next case (d) is somehow complementary to case (b). Now the two *occupied states* (in the groundstate) E_N (associated with $\mu(N)$) and E_{N-1} act as the channels for conductance. In this case only *one electron can leave the dot at a time*, for exactly the same reasons as given at (b). Evidently the availability of two channels increases the current again.
Case (e) brings us back to our starting condition, except that the dot contains one particle less, i.e. a Coulomb blockade with N-1 electrons. The current will be zero.

Figure 7.20b summarises the behaviour of the current in dependence of the gate voltage for the five cases discussed above. It clearly shows the resulting split of the Coulomb blockade peak into two. In this case we did assume the bias voltage to meet the condition $\Delta E<eV_b<2\Delta E$. Following exactly similar arguments we can show that the CB peak will split in three if $2\Delta E<eV_b<3\Delta E$, in four for biases up to $4\Delta E$, etc.

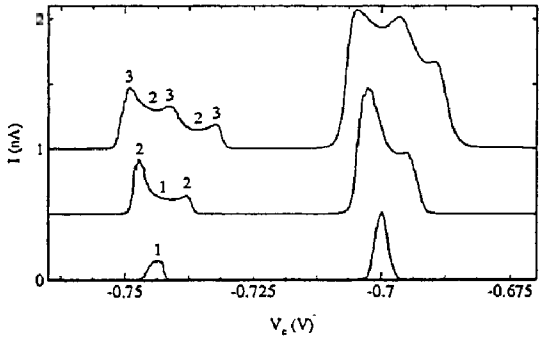


Figure 7.21 Splitted Coulomb blockade oscillations at bias voltages comparable to the 0D energy splitting. The lower trace is at small bias (100 μV), the middle trace is at 400 μV and the topmost trace is obtained at 700 μV.

In figure 7.21 we show experimental results obtained on a small lateral quantum dot of the type shown in figure 7.8. The figure contains three traces, the topmost two being offset vertically for clarity. The lowest trace shows the common small bias (V_b =100 μV) Coulomb blockade oscillations, limited in gate voltage range to two peaks. At 400 μV (the middle trace) the peaks are clearly split in doublets. Increasing the bias to 700 μV transforms each CB peak into triplets. These results clearly demonstrate the applicability of the picture discussed above. Quantitatively it allows us to derive the typical 0D level splitting of this quantum dot to be ~300 μeV.

7.2.g. *AC properties of double junction SET circuits: the single electron turnstile*

Up to now we have limited ourselves to how single electron circuits behave under stationary DC conditions for the gate voltage and the bias voltage. In the last section of this chapter we want to discuss how SET systems behave if AC voltages are applied. In particular we will see how such an AC voltage applied to the gate affects electron transport in an extension of the familiar double junction system.
From the stability diagram of Fig. 7.12 we have seen that, at a given bias voltage V, the gate voltage either leads to Coulomb blockade with a given net charge on the island, or to a finite current. An interesting additional possibility arises if the more complicated four-junction circuit of Fig. 7.22a is considered. Following similar arguments as before the associated stability diagram can be obtained. The most remarkable feature apparent from the diagram (Fig 7.22b) is the occurrence of so called *bi-stable areas*, these being the diamond-shaped figures filled with the dashed pattern. The net charge Q_n residing on the central island coupled to the gate now depends on the *history of* V_g (in figure 7.22a given as U) i.e. it shows *hysteresis*. The flat loop-shaped trajectory of V_g shown in the figure clarifies this behaviour. Starting at the far left of the V_g-cycle, the island contains zero charge (n=0). Increasing V_g makes us to traverse the hysteretic area from the left *without changing the net charge*. Only when V_g increases such that we leave the bistable area at the right, entering the n=1 range, a single charge tunnels from the -V/2 source onto the

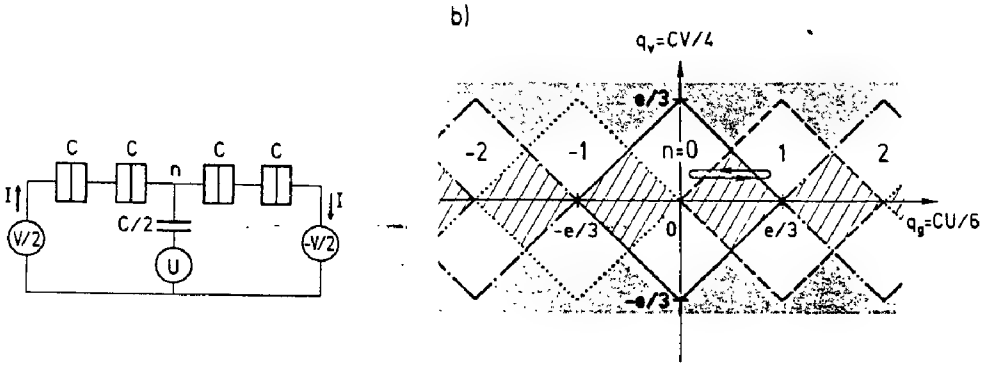


Figure 7.22. A four metallic junction circuit used as a single electron turnstile. a. represents the circuit diagram; b. shows the resulting stability diagram; note the hatched areas where the charge on the central island depends on the history of the gate voltage, i.e. it shows hysteresis.

the central island. Reducing V_g again into the bistable area *keeps the charge at* Q_n=1*e*, demonstrating the hysteretic behaviour of the circuit. Only leaving the dashed area at the left makes one electron to tunnel out of the island towards the *V*/2 source, returning the island to the *n*=0-state. Note that during this *single cycle of* V_g *a single electron* is transferred through the circuit. This is the basis for the operation of this device as a so called *single electron turnstile.*

Given this one-to-one correspondence between a single cycle and a single electron traversing the circuit, driving the gate by a potential periodic in time at a frequency *f* will result in a current given by

$$I=e*f \tag{7.26}$$

Figure 7.23 shows this remarkable effect obtained experimentally in a four-junction device made of Al-AlOx metallic junctions of a tunnel resistance R_t~350 kOhm and *C*~0.5 fF, at *T*=15 mK (B.J. Geerligs et.al, Phys. Rev. Lett. 64, 2691 (1990)). With the AC gate voltage amplitude chosen as shown in figure 7.22b clear plateaux in the *I-V* characteristics are found for frequencies *f* between 1 and 40 MHz. In addition

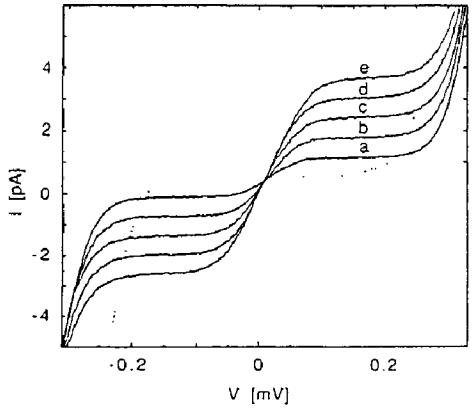


Figure 7.23 Turnstile operation of a four-junction SET device driven at 5 different frequencies. Frequency ranges from 4 MHz (curve a) to 20 MHz (curve e) in steps of 4 MHz. The dotted curve results at zero AC voltage to the gate.

to the case of no AC modulation (dotted line), fig 7.23 shows results for frequencies of 4, 8, 12, 16 and 20 MHz. Nice plateaux are found at current values in accordance with eq. (7.26) to within 0.3%. The uncertainty is determined largely by the measurement of the (very small!) currents involved and the non-ideal flatness of the plateaux.

Question: discuss a few mechanisms that may lead to deviations of the plateaux.

This result presented here, with the system being driven at frequencies far into the MHz regime, can be described fully assuming quasi static behaviour, i.e. the system just continues to be in the ground state throughout the whole AC cycle. This assumption holds as long as the *photons associated with the AC radiation* of energy E_f=*hf* do not affect the tunnel rates. More precisely, the photon energy should not exceed the energy of the photons already existing in the electron reservoirs at temperature *T*. From this we can define the crossover between the quasi-static regime and the (so called) *photon assisted* regime by *hf*~*kT*. We will not discuss this recently emerged field in any further details.

With this short note we want to close this chapter. Many aspects have not been discussed and can be found in the extensive literature covering this field. Apart from providing a much more detailed and thorough insight in this field, this also addresses the question of potential applicability of SET devices and circuits. We have seen already that the double junction circuit provides an unprecedented charge detection sensitivity, reaching at least 3 orders of magnitude lower than any existing charge detector. Work to employ this feature to study local charge fluctuations on surfaces is in progress.
From a "digital" point of view, e.g. the electron box experiment shows that a single electron can be moved in and out the island in a controlled way, raising the question whether these structures can be used as digital logic based on the concept of "one bit, one electron".
In addition the single-electron turnstile (and other variants to it) are seriously considered to be used a new and independent manner of establishing the "standard" of current. From various theoretical considerations it seems not unlikely that the intrinsic accuracy of the process should be well beyond 1 part in 10^6, or 0.0001%. Work to experimentally verify these predictions is in progress on a worldwide scale.

7.3 Charging effects in superconducting systems

In the preceding section we have discussed the effects induced by the Coulomb interaction of single particles (i.e. electrons) on the transport properties of small systems in the normal state. In this section we will concentrate on Coulomb effects in metallic systems, with the metal being in the superconducting state. As discussed in chapter 6 the superconducting state is characterised by two key properties: the relevant charge carriers in a superconductor are paired electrons called Cooper pairs, and secondly the sea of Cooper pairs is condensed into a macroscopic quantum state described by a single wavefunction. As we will show that the phase of this wavefunction and the charge due to the Cooper pairs are conjugate quantum variables it is evident that current (determined by the position-derivative of the phase) and charge are no longer independent. This has major consequences for the properties of the system.

7.3.a. *Some basic properties of Josephson junctions*

Figure 7.24 shows the typical configuration of a metallic junction made by evaporation methods, with the barrier formed by a very thin insulator like an oxide

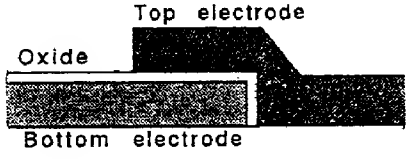


Figure 7.25. Metallic tunnel junction, with an oxide barrier.

(e.g. of the lower metal electrode). At a temperature $T<T_C$ the metal will become superconductive, resulting in the formation of Cooper pairs, i.e. pairs of electrons coupled by the (effectively) attractive force operating via electron-phonon interaction. The Cooper pairs (CP) inside the two superconducting metallic electrodes of the junction condense into a macroscopic quantum state with all CP's at the same energy E_F, the Fermi energy of the electrons. This Cooperpair condensate can be described by a wavefunction $\Psi = |\Psi| exp(i\varphi)$, with φ representing the phase of the macroscopic wavefunction. This phase will be constant throughout the metal if the current and the magnetic field are both zero. If $B \neq 0$ the phase will be affected via the vectorpotential A (see chapter 5) and it has to be replaced like

$$\varphi \rightarrow \gamma = \varphi - 2\pi \frac{2e}{h} \int A.dl \qquad (7.27)$$

Note that the relevant quantum of charge now is the Cooper pair charge $2e$, and so

the flux quantum in this case equals $\Phi_0=h/2e$, i.e. twice as small as in the case of electrons (see chapter 5).
For charging effect to be of importance we have seen that the coupling between the two bulk electrodes via the barrier has to be weak. Following the same reasoning as which leads to eq. (7.6) for the non-superconducting case, but now taking the $2e$ "unit" of charge this leads to the condition for the conductance of the barrier

$$G_t<(2e)^2/h=4e^2/h \tag{7.28}$$

From the discussion of charging effects in normal (i.e. non-superconducting) systems in section 7.2 we have seen that this "large environmental impedance" was realised by inserting an additional junction in the leads. This route is also taken in the superconducting case. So, nearly all experiments are performed in circuits containing at least two junctions. The resulting circuit, now containing an island, allows the control of the electrostatic energy by adding an external gate electrode, also completely analogues to the case of normal state junctions.

In a piece of bulk superconductor the supercurrent density j_S is related to the gradient of the phase of the wavefunction

$$j_S = const \cdot \nabla\gamma \tag{7.29}$$

Neglecting charging effects for the moment, the same relation holds for the supercurrent flowing through the junction coupling the two bulk superconductors and the phasedifference $\Delta\gamma$ between the two bulk condensate wavefunctions

$$I_S=I_0\Delta\gamma \tag{7.30}$$

Because the phase is only defined modulo 2π the relation given by (7.30) only holds for small phasedifferences. For large values it has to be replaced by

$$I_S=I_0 \sin(\Delta\gamma) \tag{7.31}$$

with I_0 denoting the critical current of the junction. This the *first* of the two famous *Josephson equations*. From the calculation based on 1D multiple coherent Andreev reflection presented in chapter 6 we found

$$I_0=(1/4\pi)\,(\Delta/e)G_t \tag{7.32a}$$

with 2Δ being the gap of the superconductor and G_t denoting the normal state conductance of the tunnel junction. More general calculations (Ambegaokar & Baratoff) have shown

$$I_0=(\pi/2)\,(\Delta/e)G_t \tag{7.32b}$$

which differs from (7.32a) only by its prefactor. If a finite difference in voltage (or electrochemical potential) ΔV is applied to the junction, B.D. Josephson showed that the phase difference starts to evolve in time at a rate given by

$$\frac{d}{dt}(\Delta\gamma)=\frac{2e}{\hbar}\Delta V=2\pi\Phi_0\Delta V \tag{7.33}$$

This expression forms the *second Josephson equation*. These two equations (7.31)

and (7.33) describe the current flow through a superconducting junction or *Josephson junction* (JJ).
Before continouing we want to introduce some simplifation in the notation. As (nearly) always only the *difference* in voltage and phase will be of importance, we will replace $\Delta\gamma$ by γ and ΔV by V.

In eqs. (7.7) and (7.8) we discussed that the rate of particle tunneling (and so the tunnel conductance) through a barrier is determined by the DOS at each side of the barrier and the transmission coefficient of the barrier as such. From this argument we can infer the general shape of the current-voltage characteristic of a single JJ, and this is shown in figure 7.26. In a. the case of zero bias voltage is shown, governed by the first Josephson equation (7.31). Here a supercurrent flows, which is linearly dependent on the difference in phase of the two wavefunctions associated with the two superconducting electrodes. Forcing $I>I_0$ will lead to a finite voltage difference due to the time evolution of the phase difference (eq. (7.33)), pictorially shown in figure 7.26b. The current which will flow is determined solely by the tunneling of

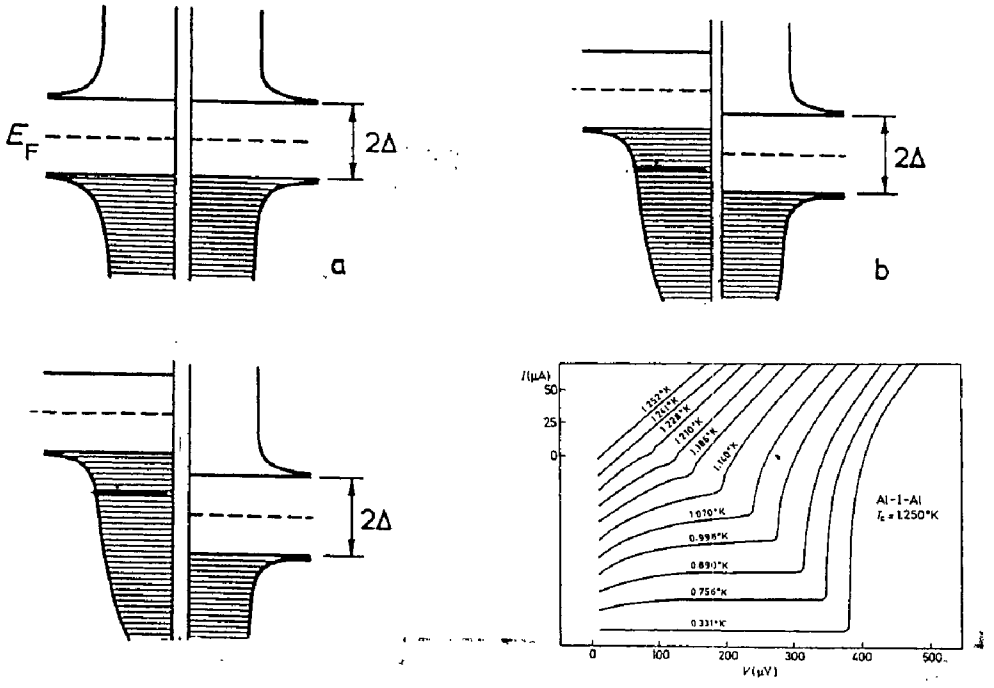


Figure 7.26. The I-V-characteristic of an S-I-S tunnel junction. a. Energy-position diagram at $V_{bias}=0$; b. at $V_{bias}<2\Delta/e$; c. at $V_{bias}>2\Delta/e$. d. I-V-curve (experimental, for Al) for a number of temperatures $T<T_c$.

single electrons (here usually called quasi-particles for reasons we will not go into). As the temperature is assumed to be very low and the number of these particles (electrons) decreases exponentially versus $2\Delta/kT$, a very small tunneling current results which is approximately linearly dependent on the bias. Effectively this leads to a very high so called *sub-gap resistance*. Increasing the bias further (figure 7.26c) to beyond the superconducting gap $2\Delta/e$ will allow particles to tunnel from occupied states below the gap in the left reservoir to empty states above it at the right. This will result in a strong increase of the current, which will depend linearly on the bias voltage. Figure 7.26d shows the resulting I-V characteristic for the full range of bias voltages.

7.3.b. *Charging effects in Josephson junctions*

In real junctions two additional elements have to be included to model the Josephson junction correctly, i.e. the capacitance C of the junction and the resistance R, this last

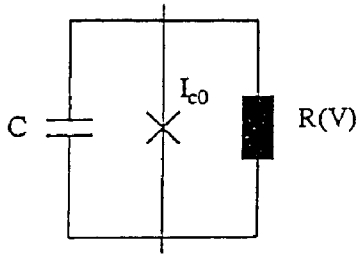


Figure 7.27. Equivalent circuit diagram of the Resistively and Capacitively Shunted Josephson (RCSJ) junction model. The "cross" represents the bare Josephson tunnel junction.

one being the sub-gap resistance mentioned above. Figure 7.27 shows the resulting circuit modelling the complete JJ. Given its components it is known as the Resistively and Capacitively Shunted Junction or RCSJ model.
The total current flowing through the circuit now simply follows as

$$I = I_s + \frac{V}{R} + C\frac{dV}{dt} \tag{7.34}$$

By using the two Josephson relations, eqs. (7.31) and (7.33), this can be straightforwardly rewritten as

$$C(\frac{\hbar}{2e})^2 \frac{d^2\gamma}{dt^2} + \frac{1}{R}(\frac{\hbar}{2e})^2 \frac{d\gamma}{dt} + \frac{\hbar}{2e}(I_0 \sin\gamma - I) = 0 \tag{7.35}$$

Inspecting this second order differential equation more carefully, immediately makes one to recognise it as the *classical equation of motion of a "particle" moving along a (generalised) 1-dimensional coordinate* γ. With this in mind the various terms in eq. (7.35) thus can be interpreted directly as:
first term: the (generalised) force resulting in the acceleration $d^2\gamma/dt^2$ of the "particle",
second term: the drag force, governed by the "velocity" $d\gamma/dt$,
last term: the force due to the gradient of the 1D "potential energy" landscape (i.e.

$dU/d\gamma$) within which the "particle" moves.
This leads to the following results

"particle mass" $$M = C(\frac{\hbar}{2e})^2 \qquad (7.36a)$$

potential energy $$U(I,\gamma) = -E_J cos(\gamma) - (\frac{\hbar}{2e})I\gamma \qquad (7.36b)$$

with E_J denoting the strength of the superconducting coupling in the junction, called the *Josephson coupling energy*, which is given by

$$E_J = \frac{\hbar}{2e} I_0 \qquad (7.37)$$

Note that, in accordance with intuition, the second (or friction) term of (7.35), shows that the movement of the particle is damped via dissipation ("heating") in the resistor R, with a large resistance (or small conductance) resulting in low damping.
Because the (co-)sinusoidal energy-phase relation given by eq. (7.36b) and shown in figure 7.28, somehow resembles an old-fashioned washboard, it is referred to as the "washboard potential" throughout the literature. So in short, one can visualise the tunneling of "electrons" through a Josephson junction with a finite capacitance as the motion in 1D "γ"-space of a "phantom particle" along a washboard-shaped potential energy landscape.

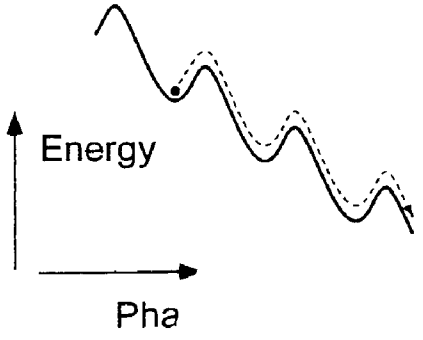


Figure 7.28. Tilted washboard potential, a model for the currentflow in a Josephson junction. The amplitude of the potential oscillations is determined by the critical current, while the tilt is governed by the total current.

The washboard is tilted by the current I flowing through the system (second term of eq. (7.36b)). At zero current the particle will sit in a minimum of the potential, i.e. in one of the valleys. As long as the tilting angle (due to the current) is such that the particle is in a local minimum of the potential, the "ball will not start to run", i.e. the phase will be stationary (apart from thermal and/or quantum fluctuations), only a supercurrent will flow and no voltage will be built up across the junction. However, if the tilt becomes so large that the gradient of the potential becomes negative everywhere (or, no local minimum will exist anymore), the particle will start to roll down the hill, leading to a time evolution of the phase ($d\gamma/dt$ becomes non-zero) and

an associated generation of a finite voltage. Consequently, also a finite charge will build up on the (capacitor of the) junction!

7.3.c. *Hamiltonian for the Josephson junction*

The preceding discussion leads us to the very important conclusion that in a JJ the charge $Q=CV$ and phase γ are *dependent variables*, coupled by the second Josephson relation $V=(\hbar/2e)d\gamma/dt$. This has far-reaching consequences, which we will discuss now.

To that purpose we have to describe the system in a more quantummechanical way. In order to do so we introduce the Hamilton operator for the system. To construct it we need the kinetic and potential energies. The potential energy is given by $U(I,\gamma)$ of eq. (7.36b). The kinetic energy immediately follows if we memorise that the particle has a mass $C(\hbar/2e)^2$ while its velocity is given by the time rate of change of the position coordinate γ, yielding

$$E_{kin} = \frac{1}{2}Mv^2 = \frac{1}{2}C(\frac{\hbar}{2e})^2(\frac{d\gamma}{dt})^2 \qquad (7.38a)$$

From the second Josephson equation (7.33) and realising that $V=Q/C$, we immediately find

$$E_{kin} = \frac{Q^2}{2C} \qquad (7.38b)$$

or, the *kinetic energy is given by the electrostatic energy of the junction due to the charge accumulating on the capacitor*.

For simplicity we assume from now on that the (subgap) resistor R is very large, leading to a negligable contribution of the resistive (or in-phase) part to the total current. With this assumption we now can "derive" the Hamiltonian by defining the total energy as the sum of the kinetic and potential energy. This yields

$$H = \frac{Q^2}{2C} - \frac{\hbar}{2e}I\gamma - E_J\cos\gamma \qquad (7.39)$$

More specifically we can define the two governing operators for position and momentum by

generalised position: γ (7.40a)

and

generalised momentum: $P = Mv \equiv M\frac{d\gamma}{dt} = CV\frac{\hbar}{2e} = Q\frac{\hbar}{2e}$ (7.40b)

7.3.d. *Energy band diagram and Bloch oscillations*

Let us concentrate first on the case $E_C >> E_J$ ($>kT$), i.e. strong charging effects; in the next subsection we will discuss the intermediate regime $E_C \sim E_J$ ($>kT$). The case $E_J >> E_C$ ($>kT$) will not be discussed here as charging effects will not be important. The properties of the resulting system can be found in textbooks on Josephson junctions in general.

In the strong charging case the total current I will be small and so H approximately only contains the *periodic term* $E_J.cos(\gamma)$ in addition to the charge term. Note that this Josephson coupling term is still small relative to the charging term $Q^2/2C$ and so in the Hamiltonian in can be taken into account as a *perturbation* term. Suppressing the very small term containing I all together, the resulting Hamiltonian *formally* resembles a very familiar case, i.e. the Hamiltonian describing the behaviour of electron waves in a crystalline lattice (see *any* textbook on Solid State physics). In that case the solutions of the Schroedinger equation describe Bloch states. The E-k relation found for these states is parabolic in case of zero periodic potential. We also know that a finite value of the periodic potential acts as a *perturbation* on the pure states, and it leads to Bragg reflections of the electron waves in real space, and to bands and gaps in the energy spectrum. The size of the gaps (linearly) depends on the strength of the periodic potential.

These standard textbook results can be "translated" immediately to our present case, reminding ourselves that

$$k = Mv/\hbar \equiv P/\hbar = Q/2e \equiv N_{CP} \qquad (7.41)$$

i.e. the solid state E-k- relations will be transposed to E-N- (or E-Q-) curves in the present case.

In the unperturbed case (E_J=0) we immediately obtain the familiar parabolic E-Q relation, as shown in figure 7.29. Adding a finite E_J will switch on the perturbation resulting in the opening of energy gaps in the spectrum at Q=(2n+1)e (or N=(n+1/2)),

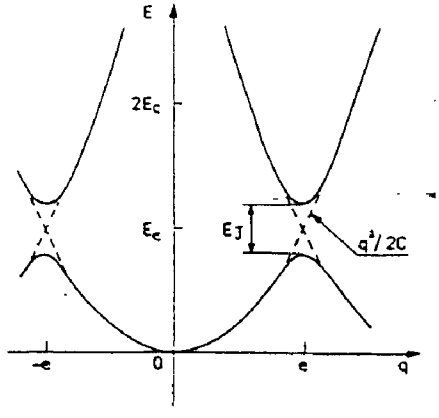


Figure 7.29: The E-Q bandstructure for a Josephson junction in the charging limit. The dotted curve is for the unperturbed case E_J=0, the solid curve is for a finite E_J.

a direct consequence of the "Bragg reflection in γ-space" by the time the charge approaches any of these particular values. Note that the size of the gaps is given by E_J. More specifically, if the "particle moving along" the Q axis (e.g.) to the right approaches the Brillouin zone boundary, Bragg reflection will start to reflect the "particle" towards $Q'=Q-2e$. This can be rephrased by saying that while time progresses the state increasingly develops into a mixture of the pure states $Q=e-\delta e$ and $Q=-e-\delta e$. Physically speaking, it represents the *tunneling of a single Cooper pair of charge 2e through the barrier of the junction*, making the charge (on one junction electrode) to change from e to $-e$.

Just as for the solid state equivalent we can choose to confine ourselves to the reduced zone scheme, i.e. $|Q|<e$. Now we have to realise ourselves that Cooper pairs all are degenerate in energy, i.e. they have no means to dissipate any excess energy ("they do not have a kinetic energy degree-of-freedom"). This means that tunneling *only* occurs if the change in energy ΔU associated with the tunneling event equals zero, i.e. *only* at $Q=e$ or $-e$. This implies zero current if the voltage across the junction is such that $|Q|<e$. This is the *Coulomb blockade* for *Cooper pairs*, with a voltage range which, for a given capacitance, is twice as large as compared to the normal-state case.

If the junction is driven by an external current source at a constant current I the Bloch state will evolve in time starting (e.g.) at $Q=0$, progressing to $Q=e$, become reflected ($=2e$ tunnels through) towards $-e$, continuing again to $Q=0$, etc. So the charge *oscillates* at a frequency given by

$$f=I/2e \qquad (7.42)$$

This is called a *Bloch oscillation*, running at half the frequency of the SET equivalent discussed in section 7.2.

For the SET case it was argued in connection with figure 7.5 that the tunneling of single electrons is a stochastic process with the Coulomb interaction acting as a regulator, in this way leading to correlation in the transfer of sequential electrons. In the present case of Cooper pair tunneling the tunneling process itself is already regulated by the phase difference between the two superconductors. The Coulomb interaction in this way just enhances the correlations in the CP tunneling.

The occurrence of Bloch oscillations have indeed been verified experimentally and we will discuss its effect on the current-voltage characteristic of an array of Josephson junctions.

In order to discuss this experiment we have to evaluate the shape of the I-V characteristic resulting from the Bloch mechanism. We assume that a constant current is fed through the array: typically this will be in the pA range, i.e. the tunnelrate equals tens of MHz. The voltage we measure in this experiment effectively is the *time-averaged or mean value* $\langle V\rangle=\langle Q\rangle/C$, as our current and voltage detectors usually will be limited to a few kHz only. Let us return to the bandstructure shown in figure 7.29. If the current source only can provide voltages below the charging value $V_C=(Q_{CP}/2)/C=e/C$ then the Coulomb blockade will suppress the flow of current. In the energy diagram this corresponds to the charge state to "sit" at a "fixed" position on the E-Q-curve at a charge between 0 and e. Only in case external influences such

as noise or thermal excitations are present these may allow the tunnelling process to occur, a process which usually follows an exponential behaviour. So, close to the origin (I,V=0) a finite voltage will sustain only an *exponentially small current*. As soon as the voltage approaches V_C the tunneling rate of Cooper pairs will increase rapidly. This will make the charge to evolve along the E-Q-curve, starting the Bloch oscillation to run as discussed before. Evidently this implies that both positive *and negative* charged states (between $-e$ and $+e$) will occur sequentially in time. This implies that the *average* value of the charge during one Bloch cycle will be (close to) zero, and so the mean voltage $\langle V \rangle$ which is measured.
In the actual experiment (L.J. Geerligs et.al, Phys. Rev. Lett. 65, 3037 (1990)) the I-V-curve (with V being the averaged value $\langle V \rangle$, as discussed!) of Al-AlOx-Al junctions is measured. The junctions are positioned in a rectangular 2D array of 190 junctions long and 60 junctions wide. Biased by a constant current source (feeding the 190 junction rows in series) the voltage over the central 95 junction rows is measured, i.e. at each side of the 95 junctions some 47 junction (rows) act as the high impedance environment. The characteristics of the junctions are R_t=35 kΩ, E_C=0.9 K, and E_C/E_J~4. Figure 7.30 shows the experimental curve measured at a temperature of

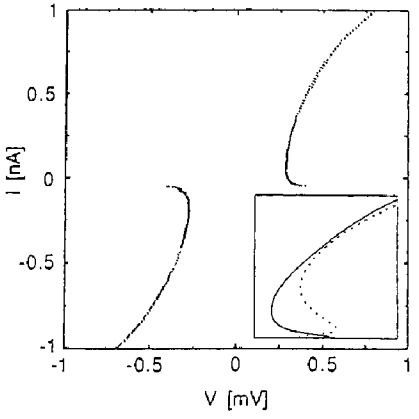


Figure 7.30. The I-V curve of a (2D array of) Josephson junctions with the charging energy E_C/E_J~4. The characteristic shape resulting from Bloch oscillations is called the Bloch nose.

~20 mK. It very clearly shows the small current branch starting from zero current and voltage, increasing to ~ 0.5 mV, followed by a reduction of the voltage at higher currents (~ 0.1 nA). Increasing the current further results in an increase of the voltage again. We have not discussed the mechanism underlying this, but it results from the finite probability for the Bloch state to continue in the *next higher energy band* of the diagram of figure 7.29. This effect is called Zener tunneling. The rather peculiar shape of the current-voltage characteristic goes by the nick-name of Bloch nose.

7.3.e. *Uncertainty relation of superconducting phase and particle number*

In subsection 7.3.c. we discussed the coupling of Q and γ, which implies quantummechanically that the associated operators P and γ are non-commuting. The immediate consequence is that the (expectation values of the) quantities γ and P can not be determined infinitely sharp *simultaneously*, but their uncertainties are

connected via Heisenberg's uncertainty relation

$$\delta\gamma.\delta P = \delta\gamma.\delta(\frac{\hbar}{2e}Q) \geq \hbar \qquad (7.43a)$$

or

$$\delta(\frac{Q}{2e}).\delta\gamma \equiv \delta N_{CP}.\delta\gamma \geq 1 \qquad (7.43b)$$

i.e. the combined uncertainty in the number of Cooper pairs N_{CP} and the phase(-difference) γ across the junction will be larger than 1.

This immediately leads to two limiting cases or regimes. Either the charge is well defined and the phase may fluctuate beyond 2π, or vice versa the phase will be rigidly determined while the charge may change beyond $2e$. This is the same as stating that either the (superconducting) Josephson coupling represented by E_J is much smaller than the charging energy E_C, or vice versa. This can be understood directly, as a (relatively) large charging energy implies a small probability for the process of adding (or removing) one charged particle from a junction electrode, i.e. δQ will be small. Now we want to concentrate on the transition region $E_J \sim E_C$. More specifically we want to study the region where the energy scale ratio E_C/E_J can be taken from <1 to >1, but not too much. From eq. (7.43b) we immediately deduce that the fluctuations in the number and phase will be of a comparable amplitude. This can be shown in a more schematic way,

$E_J < E_C$	Q (and N) well-determined	‖‖‖‖		|
		||||	Coulomb	|| Super-
		||	blockade	|||| current
$E_J > E_C$	γ well-determined	|		‖‖‖‖

To study this region more carefully one should be able to control one of the two energy scales, while keeping the other fixed. As $E_C \sim 1/C$ is determined by the geometry of the junction circuit, this is very difficult to vary *in situ* during the experiment, and so only E_J remains as a possibility to be made adjustable.

As a reminder, $E_J = (\hbar/2e)I_0$, i.e. varying E_J implies varying the critical current of the junction. Figure 7.31 shows how this is accomplished by employing two Josephson junctions arranged in a so-called SQUID (= Superconducting QUantum Interferometer Device) configuration. With a magnetic flux threading the loop that contains the two junctions (figure 7.31a) the resulting vector potential (see eq. (7.27)) will induce an additional phase difference in the condensate wavefunction around the loop. As the total phase around the loop is of course fixed at 0 (or n*2π) this implies that the phase difference across each junction will become flux dependent, and so also the critical current of the two-junction circuit will become modulated by the

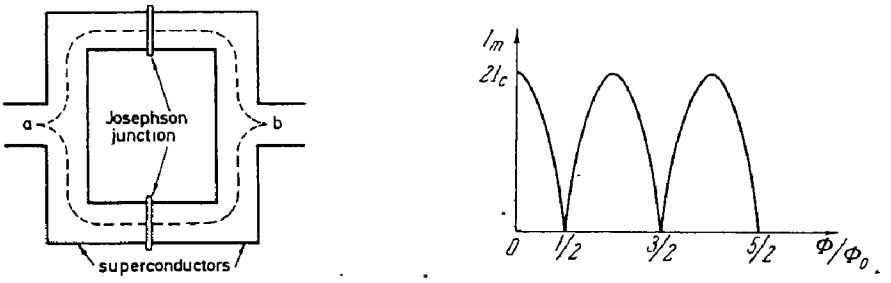


Figure 7.31. SQUID configuration employed as a device providing a tunable critical current (and E_J). a. the basic diagram of the superconducting double junction structure in a loop geometry, with a magnetic flux Φ threading the area of the loop. b. the critical current in dependence of the flux through the loop for the case of equal junctions.

phase due to the applied flux, as shown in figure 7.31b. More details can be found in textbooks on Josephson junctions and SQUIDs. If the parameters of the junctions are chosen such that

$E_C < E_J$ at Φ/Φ_0=integer
$E_C > E_J$ at Φ/Φ_0=half-integer

then the transition region $E_C \sim E_J$ can be covered properly.
In a recent experiment this so-called "Heisenberg switch" experiment has been performed (W. Elion et.al, Nature 371, 594 (1994)). Figure 7.32 shows the circuit diagram of the experiment. For this purpose the (superconducting) central island is connected to two Josephson junctions, which allows a measurement of the critical current of the structure. In addition also the "variable junction" just discussed is

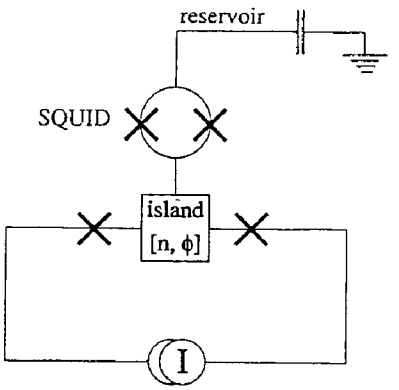


Figure 7.32. Schematic circuit diagram of the Heisenberg switch experiment. The SQUID acts as a variable E_J-coupling of the island to the Cooperpair reservoir.

connected to the island, allowing a controlled coupling of the island to a large (Cooper pair) particle reservoir. Let us concentrate on the two cases of an integer and a half integer number of fluxquanta threading the loop, i.e. the cases of strong and weak coupling to the particle reservoir. In the strong coupling case (at integer flux) the number of CP's on the island is allowed to vary rather strongly. This implies that the transport of CP's becomes more easy and so the critical current measured through the two junction connected to the current source is large. In contrast, at half-integer flux, the coupling of the island to the reservoir is weak, and so the number of CP's on the island is rather well determined. This implies that transport via the island is suppressed, or the critical current will be at its minimum. This is exactly what is found in the experiment (figure 7.33).

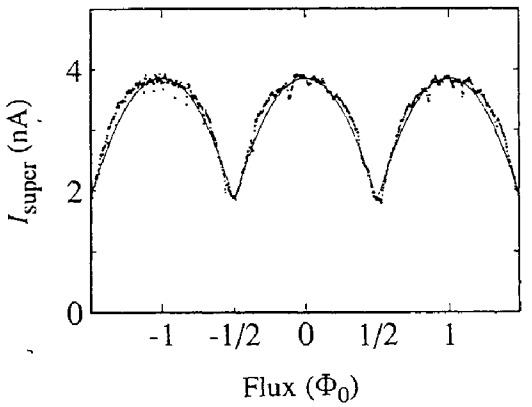


Figure 7.33.
The supercurrent in the Heisenberg circuit of figure 7.32, showing the oscillating pattern resulting from the controlled adjustment of the uncertainty in the number of Cooperpairs.

7.3.f. *Some closing remarks on charging in Josephson junctions*

In this section we only have hinted upon some of the more simple notions that are of importance in this area. Many aspects have not been discussed, mainly as these would require a much more rigid and formal (quantummechanical) approach. In particular the whole field of 2D junction arrays and the associated notion of vortices has been disregarded. Also potential applications of Josephson junction devices operating in the charging regime, e.g. as highly sensitive radiation detectors, have to be mentioned.

General references to the subject:

1. "Single Charge Tunneling", ed. H. Grabert and M.H. Devoret, NATO-ASI B: Physics 294 (Plenum, New York/London, 1992)
2. "Physics and Applications of the Josephson Effect", A. Barone and G. Paterno (Wiley, Nowe York, 1982)
3. M. Kastner, Physics Today 46, 24 (1993)
4. P. Lafarge et.al., Nature 365, 422 (1993)
5. L. P. Kouwenhoven, Science 268, 1440 (1995)

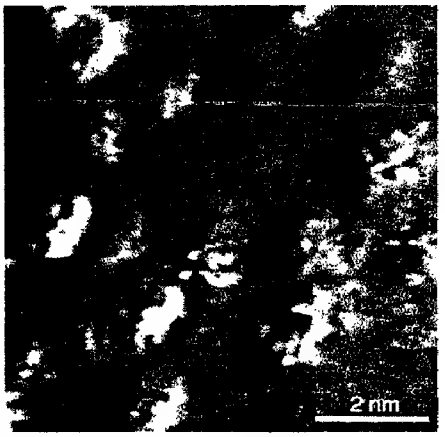
9. Thermodynamic properties of clusters

Key words: metallic clusters, energy spectra, level repulsion, shell structure, magnetisation

9.1. Introduction

In chapter 1 we have seen that single atoms may coalesce into small groups: these globules of atoms are called clusters. Under certain conditions the atoms forming the cluster may arrange in a non-random way and the resulting structure shows considerable symmetry. The degree of symmetry is governed by the underlying properties of the constituting atoms, and by the number of particles contained in the cluster.
Note that, given their structure, clusters can just as well be denoted as macromolecules.
In this chapter we will see that the (degree of) symmetry strongly determines the energy spectrum of the free conduction electrons in the cluster. As this spectrum largely determines the thermodynamics of the electronic system of the cluster, the symmetry evidently will have a strong effect on properties such as the specific heat and the magnetic and electrostatic polarisability (or magnetic and electrostatic moment) arising in a cluster.
We will discuss some of these aspects. We will limit ourselves (nearly exclusively) to systems containing free electrons, i.e. of a metallic nature. At various places one will find connections with properties of quantum dots in semiconductors as these were discussed in chapter 7 and 8.

Clusters of atoms can be made in a variety of ways, depending on the particular requirements set to the cluster. In particular the degree of coupling to any (solid) environment strongly affects the formation and so the characteristic properties of the cluster. Cluster sizes range from a few atoms to one thousand or more. Here we want to present three methods of cluster growth.
One of the "cleanest" and historically most renown methods employs an atomic beam "evaporator": the atoms of concern are evaporated and injected via a supersonic speed inert gas beam into a chamber kept at low pressure. The rapid expansion leads to strong cooling and the interaction between the atoms makes them to coalesce and condense into the cluster. As a broad range of atomic number clusters results, we require a mass separation stage behind this growth chamber in oreder to narrow the mass range of the clusters. The attractive feature of this method is that clusters are formed under nearly ideal conditions with a minimum of unwanted interactions with the environment. The less attractive aspect is the short lifetime of the cluster: once it hits the wall of the chamber it merges with many other clusters and residues, and is lost.
The second method of fabrication is of a chemical nature, using metallic colloids. In a sequence of chemical reactions, the addition of one reagent to a prepared solution first starts the formation of the bare clusters. Once these reach a particular size they rapidly start to attract long organic molecules (that are also suspended in the initial solution) which adhere to the cluster surface. This immediately inhibits the further growth of the metal cluster, resulting in a stable compound structure of metallic core plus organic cover of so-called ligands.



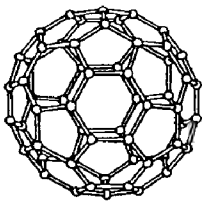


Figure 9.1. Fullerene or C_{60}-clusters on a Au surface. This first-ever image is obtained by scanning the tunneling microscopy (STM).

The stability of the compound clusters results from the repelling interaction of the ligands. The stable nature of this type of cluster allows it to be investigated for extended periods of time, however with the complicating aspect of the organic cover which seriously affects the (electronic) properties of the metal core of interest.
A third method is based on the controlled growth on a suitable substrate. In this case the choice of parameters like the type of substrate, the growth temperature and the metal vapour pressure provides ample control on the growth rate and size, and the resulting shape of the cluster (or nano crystal).
In figure 9.1 a STM image is presented of one of the most reknown macromolecules or clusters around for the last 5 years: the C_{60} fullerene or Buckeyball. Built as a (~ 2nm diameter) empty baseball of penta- and hexagons of (a total of 60) C-atoms (figure 9.1a), it show a beautiful symmetry. Figure 9.1b shows (the first) scanning tunneling microscope image (see Chapter 10) of a number of fullerenes on a Au surface (Y. Zhang et.al; see: Nature **356**, 383 (1992)).

From the preceding part it will be obvious that clusters are commonly made in rather large quantities. This is also needed as measuring the effects of interest, like the magnetisation or the heat capacity, would be very difficult to be done on a single structure. However, the control of the parameters is usually not so tight that truly single-sized clusters are formed. So, in the majority of cases cluster properties have to be derived from large ensembles of individuals, and so statistical methods are needed. The only exception is tunneling spectroscopy or (S)TM (see section 9.5).

9.2. Symmetry, energy spectra, shell structure and magic numbers

In this section we will concentrate on the mechanisms that govern the general properties of the energy spectra of electrons in a cluster of atoms. In particular the role of symmetry will be discussed.

In a confined system like an atom, molecule, quantum dot or cluster, electrons moving in the potential of the background ion(s) can not do so completely randomly because of quantum interference. In chapter 2 we introduced the phase coherence length l_φ of the travelling particle, i.e. the distance the electron can cover before it loses information about the phase it initially started with. If the characteristic dimension of the confined system is sufficiently small compared to l_φ, the electron reflecting at one boundary of the system, will (partially) revisit the area/region it came from. As it will maintain a well defined phase difference relative to its initial traversal of this area, the total resulting probability for the electron to be found at this position in space will require the summation of all partial waves *including their phase*, and so it will give rise to interference effects. In general, i.e. for a randomly chosen energy (and so wavelength) of the electron under discussion, the particle will retrace former areas with a random phase, and so summing the contributions of all scattered partial waves at a particular position in space (i.e. calculating the resulting interference) will result in a negligible total amplitude of the wave function. The interference is said to be *destructive* in nature.
If the electron waves happen to be able to retrace former paths a few times with the proper phase, the resulting interference will be *constructive* and it will lead to a "standing wave" pattern. Such a standing wave represents a 0-dimensional (0D) *eigenstate* Ψ_n of the system, and, as it only will arise at very special values of the energy of the electron, this associated energy is called the *eigenenergy* E_n of the state. So in summary

periodic motion in a confined system <==> eigenstates and eigenenergies

Note that this reasoning is at the basis of the formation of the energy spectra of atoms and molecules, but also at that of the frequency spectra of musical instruments, bridges, etc. provided we replace "electron wave" by "elastic displacement".

Before proceeding to see what the effect of the shape of the confined system is on these states, let us first consider the role of the *phasecoherence*. In an atom or molecule in the ground state the electrons are residing in (closed) atomic or molecular orbitals, and they will continue to do so indefinitely, i.e. the phase coherence time is very long. In a quantum dot and a cluster the situation may be different. The electrons inside the system may couple to other states, e.g. the phonons residing in the ion lattice forming the background potential. For the phonon example the coupling will induce *inelastic scattering* events, which modify the energy of the electron and so the phase. The time the electron can stay in the pure state will now become finite, and this finite phase coherence time τ_φ immediately leads to a finite energy width $\delta E \approx h / \tau_\varphi$ associated with the state, determined by the uncertainty relation.
A similar effect on the energy states results if the confinement is not complete, i.e. if the electrons can "leak away" from the confining volume. This is the case if the cluster or dot is connected to leads, e.g. to allow transport measurements (see chapter 7). Now the coupling of the electron wave to the leads will reduce the fraction of the wave that ultimately can retrace its previous path and so it will lead to a finite lifetime of the state and a finite energy width.
It is rather instructive to "derive" a rough estimate of the relationship between this finite width, and the leakage rate and 0D energy state splitting $\Delta E = E_{n+1} - E_n$. Assume a system

of characteristic size L, connected to the external world via leads with a transmission probability T (or a total conductance $T.e^2/h$). The velocity of the state of energy E_n will be

$$v_n = \frac{1}{\hbar}\frac{dE}{dk} \sim \frac{1}{\hbar}\frac{\Delta E}{\Delta k} = \frac{1}{\hbar}\frac{E_{n+1}-E_n}{k_{n+1}-k_n} \tag{9.1a}$$

For the two states $\{n+1\}$ and $\{n\}$ the difference in k-vector will be typically $\Delta k \sim \pi/L$. The typical time to perform one roundtrip in the system will be $\tau_{per} \sim L/v_n$ and the time it takes for the electron wave to leak away will be approximately

$$\tau_{leak} \sim \tau_{per} / T \tag{9.1b}$$

This leakage time now effectively acts as the phase coherence time. Taking these results together yields for the energy width due to coupling

$$\delta E \approx \frac{h}{\tau_{leak}} \sim \frac{2\pi T}{L}\frac{\Delta E}{\Delta k} \sim T\Delta E \tag{9.2}$$

This simple expression provides a rough estimate for the effect of wave leakage.

Now that we understand the effect of inelastic contributions, what will be the effect of *elastic* events? Elastic events only affect the path the wave will take, and so a different energy and wavelength will be required to set up the resonance or standing wave condition. Stated differently, elastic scattering effectively work out in *exactly* the same way as the reflection at the boundary of the system: just as the boundary conditions determine the eigenfunctions, so do elastic scatterers as these lead to *changes of the shape of the spatial potential* of the confined system.
To evaluate the effect on the energy spectrum resulting from a change of the confining potential we simply can use the "particle-in-the-box", taking a 3D rectangular shape. Limiting ourselves to the simple independent electron case, the three directions are mutually decoupled (or orthogonal) and the eigenenergies are a simple sum of the three contributions resulting from the k-quantisation in the three directions in space. Figure 9.2 shows the energies for two symmetry cases. In figure 9.2a a fully symmetric cubic box is chosen ($L_x=L_y=L_z$) while figure 9.2b shows the case $L_x<L_y<L_z$. The most obvious conclusion we can draw from this simple example is that the highly symmetric case shows a considerably more simple and less dense spectrum. Concerning the density, this of course is only an appearent effect, as there are not fewer states available in the symmetric system below a certain energy, but each state is strongly degenerate due to the large degree of symmetry. This is a very general feature of energy spectra for independent particles: the larger the degree of symmetry, the larger the typical level splitting.
This immediately allows us to conclude on what will happen if disorder (like elastic scatterers) is introduced in an initially regular system. The disorder evidently reduces the symmetry and so it will lead to a more complicated and denser spectrum, with less degeneracy per individual state. The ultimate case will be that all (orbital or path)

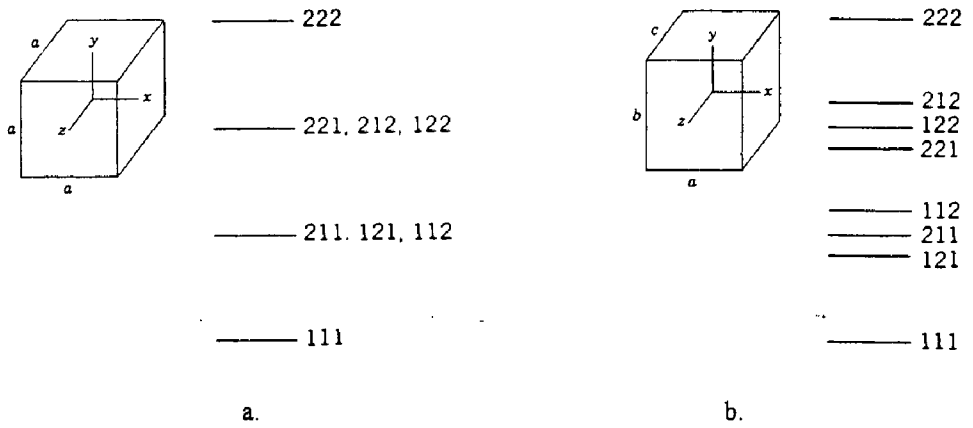


Figure 9.2. Energy spectra for a particle in a rectangular box with hard walls and of (x,y,z)-sizes a, b and c; numbers next to each eigenenergy indicet the associated quantum numbers in x, y and z. a cubic symmetry, a=b=c; b block symmetry, a, b, c all different.

degeneracies will become lifted. In the simple single electron picture this will result in two electrons per state due to spin at zero magnetic field, becoming lowered to a single electron per state if also the spin degeneracy is lifted (e.g., by a magnetic field).

9.3 Shell structures and magic numbers

As discussed very briefly in chapter 1 atoms in clusters may arrange in such a way that a clear preference for certain group sizes results. This becomes visible in mass spectroscopic data which show an increased abundance of clusters containing particular numbers of atoms, and these numbers are denoted magic. To understand this phenomenon, which has its counterpart in the build-up sequence of the electron clouds around an atom as well as on the arrangement of nucleons inside the atomic core, a rather simple approach can be taken. Assume that at the formation of the cluster each atom provides one electron to the "conduction electron sea", leaving a positive background charge of the ions onto which the electrons are bound.

Figure 9.3 shows the calulated energy spectra for particles in two different confining potentials (W.A. de Heer, Rev. Mod. Phys. 65, 611 (1993)). The leftmost picture of figure 9.3a. is for a 3D parabolic potential leading to harmonic oscillator behaviour. The single particle levels are nicely equidistant in energy, numbered by the single quantum number n=0, 1, 2, 3, ... noted just above each level. Also, in parentheses, the degree of degeneracy is given due to the symmetry of the system, which is 2-fold at n=0, 6-fold at n=1, etc. Summing these degeneracies one finds the series of numbers 2, 8, 20, 40, 70, etc...

Let us see next what happens if the rectangular confining potential is employed. The rightmost picture in figure 9.3a shows the resulting level structure for a 3D square well

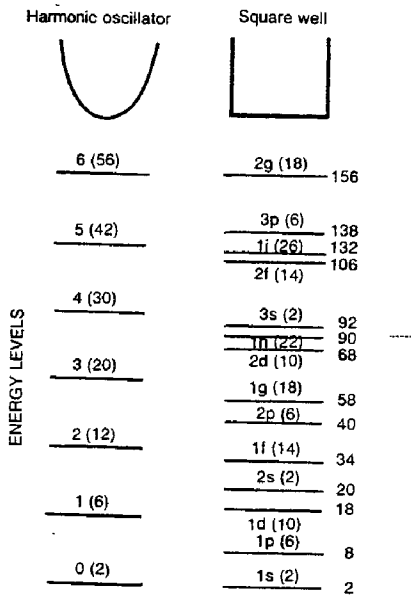


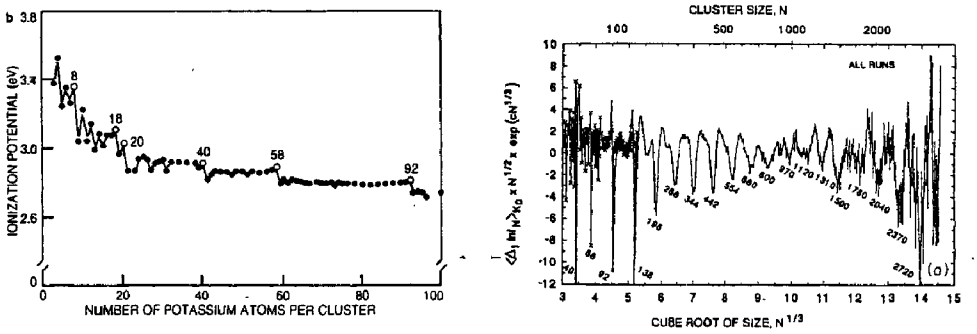
Figure 9.3. Energy level spectra for two different confining potentials. At the left a 3D parabolic potential is chosen. At right a 3D hard wall square well is taken.

confinement. The spectrum becomes much more complicated, and requires two quantum numbers $\{n,l\}$ to be enumerated: in the figure these are shown directly above each individual level. Again the degeneracy of each level is given in parentheses and the total number of particles that can be put in the given number of levels is given at the right, as 2, 8, 18, 20, 34, etc. Note the clustering of the levels, with larger "gaps" above particle numbers 2, 8, 20, 34, 40, 58, 92, ...: the *magic number* sequence.

Now, what do these levels imply? The total energy of the system can be evaluated simply as

$$E_{tot} = \sum_n g_n E_n \qquad (9.3)$$

with g_n denoting the degeneracy of the n-th level E_n. Evidently the energy differences at the addition of one atom (+electrons) after the other will show a step whenever the n-th level becomes completely filled and the next (i.e. $(n+1)$-st level) has to become occupied. Such an extra large difference in energy will occur in going from the filling numbers 2->3, 8->9, 20->21, 40->41, etc. With this fact in mind it is not difficult to see that a cluster of size 2, 8, 20, 40, ... will be significantly more stable than those at 3, 9, 21, 41, ... Rephrasing this "stability" argument one can say that it implies a stronger binding of the particle to the cluster, and so a larger energy will be required to ionise the system, whenever the cluster size matches one of the magic numbers. So, we anticipate that the ionisation energy should show a step-wise increase at each magic number size. Similarly we understand the preference for clusters to grow in sizes containing magic numbers of particles.



a. b.

Figure 9.4. Experimental results showing the increased stability of clusters of magic number size. a. the ionisation energy of K clusters of different size, measured in the UV range; b. the relative abundance of Na cluster masses, obtained from a mass spectrometer

Figure 9.4 shows that indeed this stability argument shows up experimentally. In figure 9.4a (M.L. Cohen et.al, J. Chem. Phys. 91, 3141 (1987)) the ionisation energies of an ensemble of K clusters, fabricated in an atomic beam evaporator system, is shown. A clear structure is manifest at the magic numbers like 8, 18/20, 40, 58 and 92. In figure 9.4b the picture shown in chapter 1 is taken again (J. Pederson et.al, Nature 353, 733 (1991)), displaying very clear evidence for an increased abundance of the magic number cluster sizes in the mass spectrum.

9.4 Energy level repulsion and statistics in quantum-confined systems

The question we want to address now is: what degree of (in-)dependence is there between the eigenstates and eigenenergies of systems with *elastic* scattering. One obvious effect of such scattering will be that the pure states of the "clean" (i.e. scattering-free) system will become mixed whenever the elastic scattering is introduced. To appreciate how this affects the energy level structure we assume that the mixing is not too strong, which allows the use of perturbation methods. Assume that the pure, clean system has eigenstates $\Psi_{i,0}$ and eigenenergies $E_{i,0}$. The scattering can be formally described by a Hamiltonian H_s. It effectively mixes the pure, unperturbed states $\Psi_{j,0}$ to form the perturbed states Ψ_i via the matrix elements

$$H_{ji} = <\Psi_{j,0} \mid H_s \mid \Psi_{i,0}> \qquad (9.4)$$

Using second order perturbation theory we immediately obtain the new, shifted eigenenergy of the perturbed i-th. eigenstate

$$E_i = E_{i,0} + H_{ii} + \sum_{j \neq i} \frac{|H_{ji}|}{E_{i,0} - E_{j,0}} \qquad (9.5)$$

From this expression we see that the largest contribution to the shift of an eigenstate is due to states that mix apreciably (i.e. H_{ji} relatively large) and with a small energy difference while unperturbed (i.e. $E_{i,0}$-$E_{j,0}$ small). Rephrasing this, it means that the *largest shifts will be found whenever the unperturbed eigenenergies of two states are (nearly) degenerate*. This implies that levels tend to *prevent to cross one another*, a phenomena called *level repulsion*. Figure 9.5 shows in a very schematic way how the perturbation modifies the unperturbed energy level diagram. At the position of the original crossings the level repulsion leads to energy gaps.

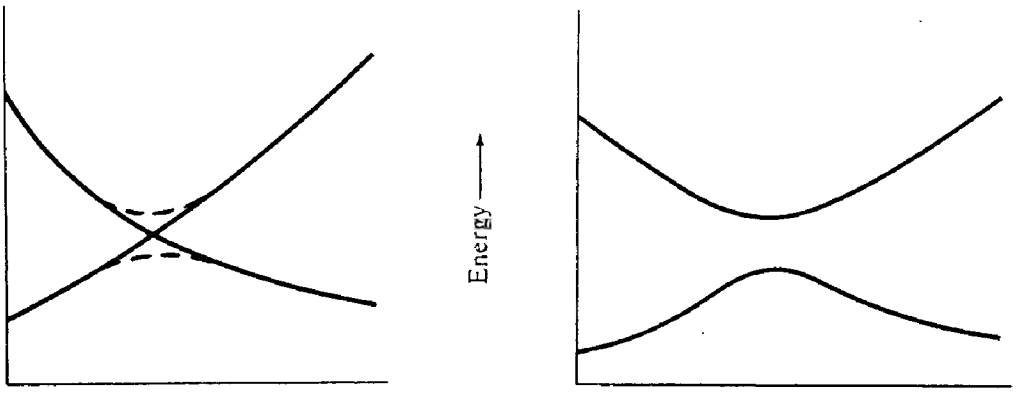


Figure 9.5. Level repulsion resulting from perturbation action. a. the unperturbed groups of energy levels and how the perturbation opens an energy gap at the position of the crossing in the unperturbed state; b. the resulting level diagram with the gaps.

This phenomenon is not as unfamiliar as it may seem at first sight: it is this mechanism which opens the gaps in the E-k spectrum in the solid state due to Bragg reflection of the electron waves at the periodic potential of the lattice (i.e., eleastic scattering); we have also seen it to occur in the (calculated) level diagram for the persistent currents in chapter 5 for a ring with elastic scatterers. Note again that strong scattering leads to strong level repulsion and so to narrow bands and wide gaps, a conclusion we have formulated in chapter 5 before.

From the preceding discussion we can conclude that the phenomenon of level repulsion no longer allows energy levels to behave independently: *level repulsion leads to correlation between individual energy levels*. This interdependence evidently should have

consequences for the statistical distribution of levels in (complex) spectra. Wigner (1951) and Dyson (1962) were the first that became aware of this consequence. Let us concentrate on this a little more.

Assume that we have a relatively large system. The calculation of the exact energy spectrum will be very complicated (as all interactions have to be included) and is commonly far beyond the capacity of present day computers. However, for a number of properties of clusters it suffices to evaluate the *statistical properties* of the energy levels and the resulting density-of-states.

To that purpose we define the probability $P(\Delta)$ which characterises the level density fluctuations across the total spectrum. More precisely, $P(\Delta)d\Delta$ is the probability to find the nearest (or neighbour) level at an energy $(\Delta,\Delta+d\Delta)$ away. This can be evaluated over the energy range of the complete spectrum or over a part of it (but it should contain a reasonably large number of levels to guarantee statistical significance). For the following discussion it is convenient to choose the average 0D level spacing or splitting δ as the "unit of energy difference"; typically we will have $\delta=g_s(E_F/N)$ with N being the number of electrons in the cluster (see chapter 2 and 7). If no level repulsion would occur the levels will be distributed approximately randomly relative to each other and so they will show a Poisson-distribution

$$P_0(\Delta)=\frac{1}{\delta}exp(\frac{-\Delta}{\delta}) \qquad (9.6)$$

Figure 9.6. (solid curve) shows this probability distribution; note the large probability to find levels in close proximity from one another, and the rather small chance for large level-to-level energy differences.

However, if the system *does* show repulsion, it is evident that we should have a *small* likelihood for the occurrence of levels being *closely spaced*. This is shown (only schetchy!) in the same figure by the dashed curve, where at small energy spacings the probability is small. Note that, in agreement with intuition, the curve reaches its maximum at approximately the average level spacing, i.e. at $\Delta\sim\delta$.

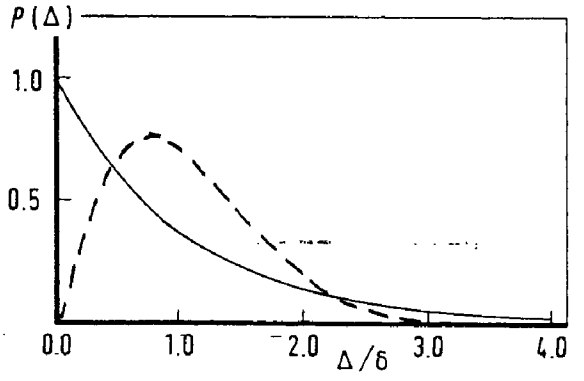
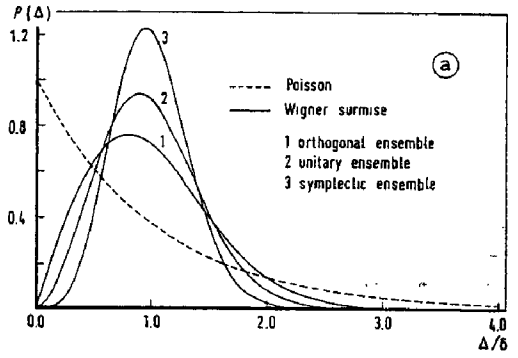


Figure 9.6. The probability distribution of the energy level spacings Δ for a random (Poisson) case (solid curve) and for the case of level repulsion (dashed).



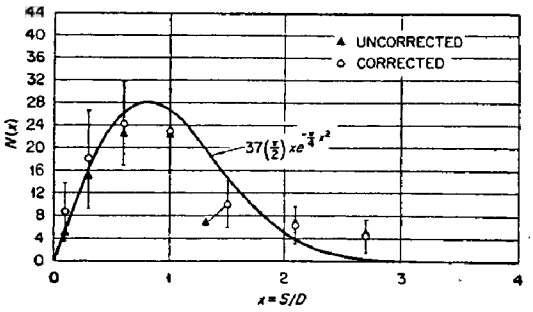
Figuur 9.7. The nearest-level probability distribution for the three classes of symmetry (1 orthogonal, 2. unitary, 3. symplectic) and for the independent level case (Poisson, dashed)

The exact expression for the probability distribution function depends on the *symmetry* of the system, and so on the Hamiltonian describing it. Without discussing the details behind it, we just present in figure 9.7 the three probability distributions for three distinct classes of systems characterised by a particular symmetry. These are:

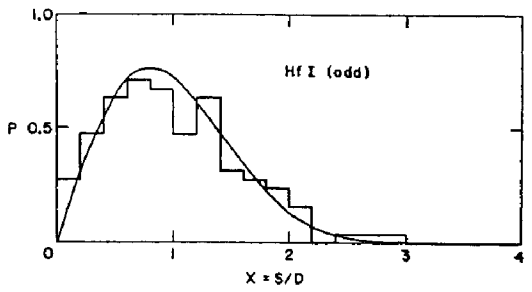
Symmetry	Allowed transformations	Distribution function	
time invers. + int. spin time invers. + space rot.	----orthogonal	$P_1(\Delta) = c_1 \frac{\Delta}{\delta} exp(-b_1(\frac{\Delta}{\delta})^2)$	(9.7a)
no time-inversion	unitary	$P_2(\Delta) = c_2(\frac{\Delta}{\delta})^2 exp(-b_2(\frac{\Delta}{\delta})^2)$	(9.7b)
time inversion	symplectic	$P_3(\Delta) = c_3(\frac{\Delta}{\delta})^4 exp(-b_3(\frac{\Delta}{\delta})^2)$	(9.7c)

Note that from the determination of the level distribution some of the underlying symmetry of the system should become apearent. On the other hand the differences between the curves is not very large, so to reach a statistical significant distinction requires the measurement of a large number of level-differences!

As we have been stressing, these level distribution properties are of a very general character, allowing it to be applied to any "large" *coherent* system of particles described by a Hamiltonian of certain symmetry. To underline the universality of this approach we show three examples: one from level statistics of the spectrum of a nucleus, the second for the case of an atomic (i.e. the electronic cloud) spectrum; the third refers to resonant frequencies in microwave cavity. Figure 9.8a shows the result obtained on the level



a.



b.

Figure 9.8. Energy level distributions of complex many-particle systems. a. Nuclear level distribution of four different zero-spin U nuclei. b. Atomic level distribution of the odd parity electronic levels of Hf

distribution of four zero-spin nuclei of U^{234} and U^{236} (J. A. Harvey et.al, Phys. Rev. 109, 471 (1958)). Despite the rather large spread of the experimental results, the similarity to the orthogonal distribution $P_1(\Delta)$ is striking! (note: S/D in this figure corresponds to our notation Δ/δ).

In figure 9.8b the results for the level spacing statistics of the electronic (odd parity levels) spectrum of atomic Hf is shown (N. Rosenzweig et.al., Phys. Rev. 120, 1698 (1960)). Again, without going through all the details it is quite evident that a good qualitative agreement is obtained for the shape of this distribution.

With the aim to stress the universality of the level repulsion phenomenon, a completely different system is taken as the third exemple (H.J. Stockmann et.al., Phys. Rev. Lett. 64, 2215 (1990)). It comprises the distribution of resonant frequencies (in the microwave region) of a closed (2D) cavity (figure 9.9a). Each negative going peak is associated with one eigenmode of the structur. With typical sizes much larger than the longest wavelength

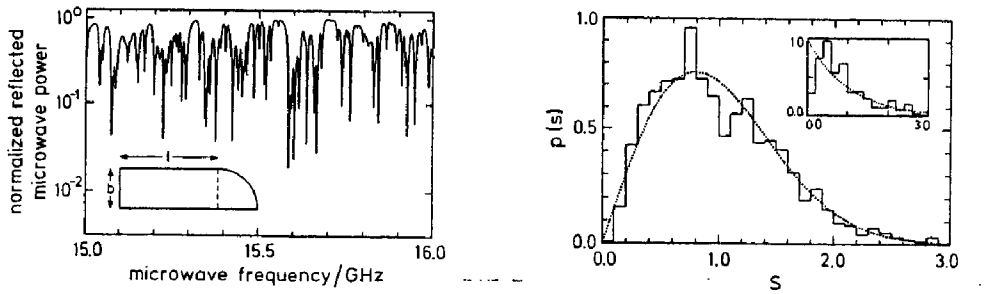


Figure 9.9. Resonance frequency distribution of a so-called stadium billiard cavity. a. The spectrum of the reflection coefficient of the cavity (the inset shows the shape of the cavity). b. Histogram of the energy level spacings.

used in the measurement, the statistics of the resonance frequencies (= eigenfrequencies associated with the standing wave modes) is determined (Fig 9.9b). The quite striking level-repulsion shape is obvious.

Before continuing with the next section one remark should be made. From the preceding discussion it is now clear that a variety of properties in small *coherent* systems is governed by the electronic energy level distribution and their occupancy. This holds particularly for those quantities that are defined by the ground state of the system. However, also effects involving the occupation of low-lying excited states, like electronic transport, are determined by the properties of the electronic energy spectrum. Clearly for all these aspects not the actual values of the *individual* matrix elements involved in the description are of importance but only their *mean value and statistics*, including the symmetry. This is on the bases of a widely applied method called Random Matrix Theory (RMT). It allows the calculation of mean values of a range of quantities (like the magnetic moment or the conductance) of ensembles of systems, and so of the typical value of that particular quantity of one individual member of the ensemble. By the same token it provides quantitative information on the statistics of the particular quantity as well.

9.5 Thermodynamic quantities; parity effects

Quantities such as the magnetic susceptibility and moment, and the heat capacity (or specific heat) are determined by the thermodynamical properties of the system. To evaluate these quantities it suffices to calculate the free energy F and take the appropriate derivative. In this section we will describe how this problem should be approached.

The mentioned quantities are given by

$$M = -\frac{\partial F}{\partial B} \tag{9.8a}$$

$$X = \frac{1}{V}\frac{\partial M}{\partial B} = -\frac{\mu_0}{V}\frac{\partial^2 F}{\partial B^2} \tag{9.8b}$$

$$C_V = \frac{\partial F}{\partial T} \tag{9.8c}$$

with V denoting the volume of the cluster. Note that eq. (9.8a) has been used before: to calculate the orbital magnetic moment due to the circulating persistent current in a metal ring (see chapter 5) it was applied to the energy levels resulting from the electron waves encircling the loop in a phase-coherent way. To obtain the total moment we had to sum the contributions of all the individual filled states, i.e. for all the energy bands below the Fermi energy of the ring. Note that the moment evaluated in this way is due to the *orbital motion* of the conduction electrons: no effect of the spin was taken into account. One important consequence of the summation over all the individual electron states was the large degree of *cancellation* and the effect resulting from the *parity* of the electron number: the addition of one electron to the ring completely alters the total circulating current (and so the magnetic moment), even including a change of sign.
In the present case of the cluster it is evident that a similar orbital contribution may arise, although the different sizes of the various orbits that are allowed in the cluster as compared to the rather restricted paths in a ring configuration will introduce modifications in terms of the magnetic field dependence. Given the considerable similarity to the persistent current case we will not rediscuss it here for clusters, and concentrate on the effect of the spin of the electrons. It should be noted that the total magnetic moment (as well as other thermodynamic quantities) is a sum of a variety of contributions, such as due to the nuclei and the electrons bound to the constituting atoms (ions) of the cluster.

From statistical mechanics we know that in order to calculate the free energy $F(N,T,B)$, with its dependence on the electron number N, the temperature T and the magnetic field B, we have to evaluate the partition function (or "toestandssom")

$$Z(N,T,B) = 1 + \sum_j exp(-\sum_i n_{ij}E_j / kT) \tag{9.9}$$

with E_j denoting the eigenenergies with degeneracy (or occupation) n_{ij}. The free energy follows immediately form the partition function

$$F = -kT \ln Z \tag{9.10}$$

Intuitively one may feel that the strongest effects due to the confinement will occur whenever the temperature is low, i.e., such that the average level spacing $\delta = g_s(E_F/N) > \sim kT$. This implies that only the states closely to the Fermi energy will become

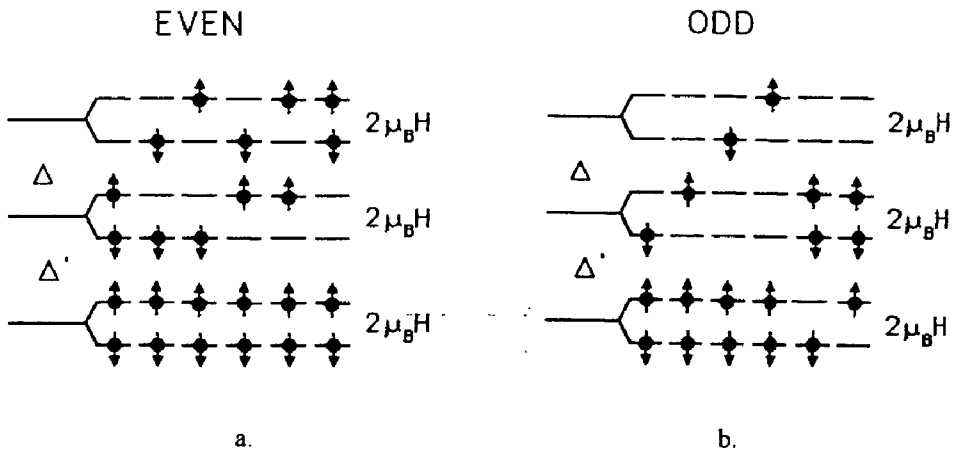


Figure 9.10 The electronic level structure diagrams for the two cases of clusters with an even (a., left) and odd (b. right) number of electrons. For finite magnetic field (right) the lowest five excited states are given from left to right, in addition to the ground state (leftmost configuration)

affected by the finite temperature. In order to demonstrate the approach of such a problem, we simplify matters as much as possible and assume only three levels to be of importance (figure 9.10), with only the lowest two being occupied at zero temperature. At zero magnetic field the levels, at distances Δ' and Δ ($\sim\delta$), are shown at the left of the figure, assuming an even (odd) number of electrons (and so spins) for the left (right) part of it. With a small magnetic field B applied each level will split symmetrically around the zero field level, at a distance + and - $(1/2)g\mu_B B$ (g is the Landee factor which is ~ 1.00 for free electrons) for up-spin and down-spin respectively. We assume the magnetic field such that $g\mu_B B < \Delta', \Delta$. The leftmost configuration is the ground state with 4 spins for the even case and 3 for the odd case. The other five configurations (going from left to right in the two figures a and b) represent increasingly higher excited states.
From figure 9.10 we can derive qualitatively that the electron number parity (even or odd) has a dramatic effect on the magnetic susceptibility. In the *even* case at a small magnetic field and temperature, $g\mu_B B$, $kT \ll \delta$, the two spin states of the central level will be occupied, but no higher states will become excited. Consequently no spin flips can be induced by the magnetic field and so the magnetic moment (as well as the susceptibility) will be (exponentially) small. For the *odd* case at small field $g\mu_B B < kT \ll \delta$ the electron is trapped in a two-fold spin degenerate state where it very easily can flip from the up- to the down-state. This implies a strong polarisation of the spins for a very small field already, i.e. a large susceptibility. As the electron behaves more or less as a free electron it is likely to follow the standard Curie law: $X = \mu_B^2/kT$.

Now let us consider an ensemble of clusters. Taking the centre level as the zero of energy, the partition function eq. (9.9) can be calculated straightforwardly. The result thus

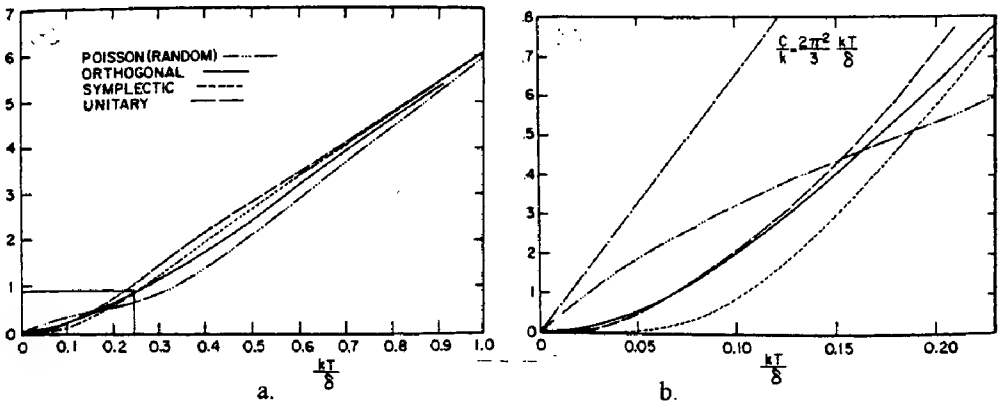


Figure 9.11. The cluster heat capacity as a function of temperature. An average of half odd and half even clusters is taken. b Expansion of a close to zero temperature.

obtained holds for *a single particular cluster* containing N electrons and with two *particular* level splittings Δ' and Δ. Now we have to realise that, experimentally, one commonly investigates a sample containing a large ensemble of clusters with a variety of shapes, and some distribution in number of electrons per cluster. This implies the need to average over different *level distributions,* in accordance with the symmetry rules presented above, i.e. by selecting the appropriate level distribution function $P_m(\Delta)$ from eqs. (9.7) and integrating over all possible level configurations weighted by the distribution function. This yields the level-averaged value $\langle \ln Z \rangle$. Now we can take the appropriate derivative to obtain the magnetic moment etc., following eq. (9.8). The details of the calculation can be found in Halperin (1986), and we only show the results. For the specific heat one finds a power-law behaviour versus temperature, which does not differ in its exponent for the odd and even case. They do, however, differ for the different symmetries. In figure 9.11 the specific heat (eq. 9.8c) is given for the four different symmetries,
calculated for the case that half of the clusters is odd and half is even. The lower part (b) of the figure is an expansion of the top part (a), showing the behaviour at small T in more detail. Notice the strong deviation form the bulk behaviour $C_{bulk}=2(\pi^2/3)k^2T/\delta$ (see e.g. Kittel, Intr. to Solid State Physics)
In exactly the same way also the magnetic moment and susceptibility can be calculated. Here the behaviour shows a significant difference between even and odd, as we have seen already in our qualitative discussion. For the *odd* case (figure 9.12) all distributions show a clear $1/T$-behaviour, in accordance with our previous arguments, and only marginal differences are apparent for the different symmetries. The *even* case also follows our intuitive reasoning, showing a decrease with decreasing temperature.
Looking to the odd case in more detail we have to note that our derivation only holds if $g\mu_B B<kT$, as this guarantees the up- and down-state to be nearly energetically degenerate. In practical systems we know that a number of other mechanism (i.e., other than the free electron spins as discussed here) may contribute to the magnetic state of the cluster (nuclear, ionic, orbital moments). The magnetic field resulting from these contributions

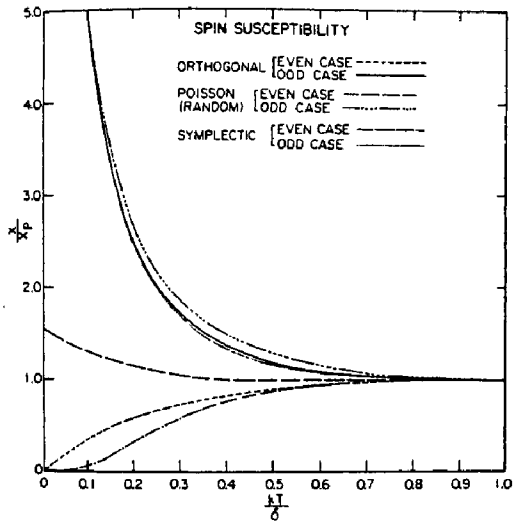


Figure 9.12. Cluster magnetic (spin-)susceptibility for three symmetries. Note that all odd parity cases follow very similar 1/T-curves. In contrast the even case shows a clear suppression for $kT<\delta$, except for the random Poisson distribution.

implies this condition will fail to hold at ever decreasing temperature. This will imply a saturation of this odd spin contribution.

Now that we have discussed the theory at some length we want to present some more experimental results. Two recent experiments will be described demonstrating the 0D levels in small metallic particles, and one experiment showing the effect of the parity on the magnetic susceptibility of an ensemble of clusters.

The first experiment to be discussed has already been presented in the Introduction, chapter 1 (M.E. Lin et.al, Phys. Rev. Lett. 67, 477 (1991)). Figure 9.13 shows the set-up. The basic idea behind the experiment is to employ the effect of field emission of electrons from an area of large electric field strength to extract electrons from the cluster. These emitted electrons are collected in an energy analyser, allowing their energy to be determined with high resolution.
In fig 9.13a the set-up is shown schematically. The cluster, attached to a *very sharp needle*, is held at zero potential. A nearby screen, with a small hole in its centre, is put at a large positive voltage. The sharpness of the tip leads to a very large electric field in the vicinity of the tip, resulting in *field-emission* of electrons from the tip. If a small object is located close to the sharpest area of the tip, the field will be largest there and so electrons from this object will be emitted most easily. This is shown in fig 9.13a. In fig 9.13b an energy diagram for the tip is displayed, with the cluster modelled as a potential well and the bound states of the electrons shown inside the well. Electrons emitted from these bound states (via tunnelling through the triangular barrier at the right) and arriving at the screen (or the analyser) will have a *total energy* given by the sum of (e times) the acceleration voltage *and the bound state eigenenergy*;

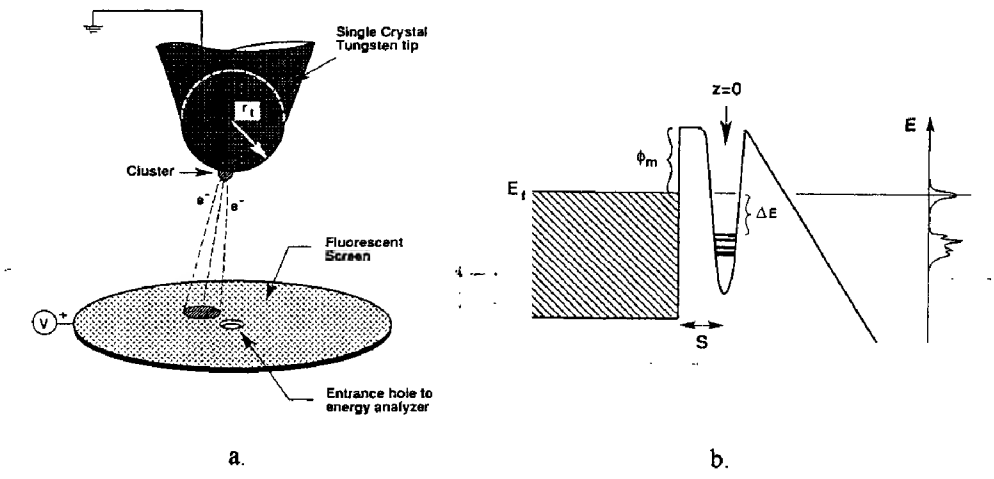


Figure 9.13. a. Schematic diagram showing the cluster attached to the field-emitting tip. b. Energy landscape showing the bound states of the cluster, and the triangular barrier separating the eigenstates from the vacuum towards the fluorescent screen.

note that this total energy will be *less* than the accelerator voltage. In the analyser, located behind the hole in the screen, each incoming electron is counted and its energy is measured

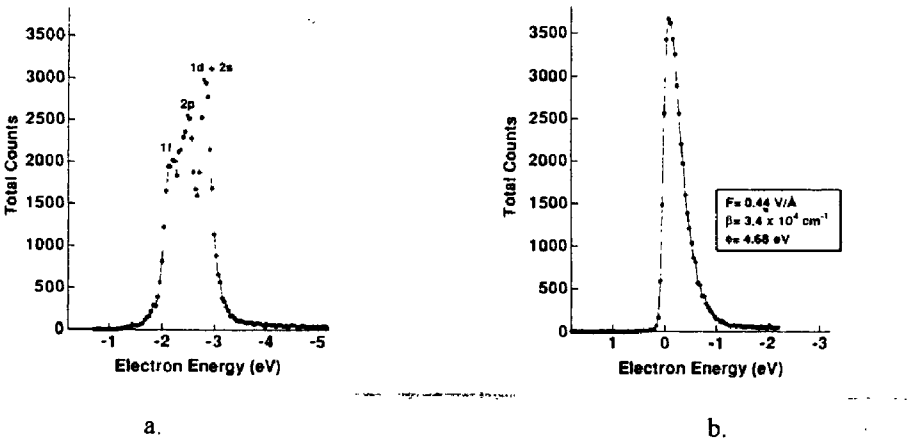


Figure 9.14. a. The energy spectrum from the field-emitted electrons from a Au cluster of ~ 1 nm diameter; the total applied voltage difference is 2550 V, stressing the need for very high energy resolution. b. For comparison the spectrum for electrons emitted from a clean tip; note the featureless peak.

leading to a spectrum of the number-of-electrons <-> energy. As the electrons from the bound states have different total energies they should show up in the spectrum as different peaks. This is exactly what is seen in fig. 9.14a and b. In particular 9.14b shows what is found for a clean tip: no particular structure in the spectrum is seen. In contrast fig. 9.14a shows the spectral distribution for a Au cluster of ~ 1nm diameter attached to the tip: a clear splitted-peak-shaped distribution (at an energy *less than the accelerator* voltage, as anticipated) is evident. This result demonstrates the bound state eigenenergies of a metallic cluster.
Before closing the discussion of this experiment we should mention one complicating factor. As discussed in chapter 7 the small capacitance C associated with the small size of the particle will result in an appreciable Coulomb energy $\sim e^2/2C$ required to add or remove an electron to or from the cluster. It is very likely that in this experiment charging energy plays a significant role. To what extent it affects the results can not be determined immediately and so the conclusions from this experiment should be treated with some caution!

The second experiment is more recent (D.C. Ralph et.al., Phys. Rev. Lett. 74, 3241 (1995)) and employs an Al island embedded in an insulating medium. By voltage bias

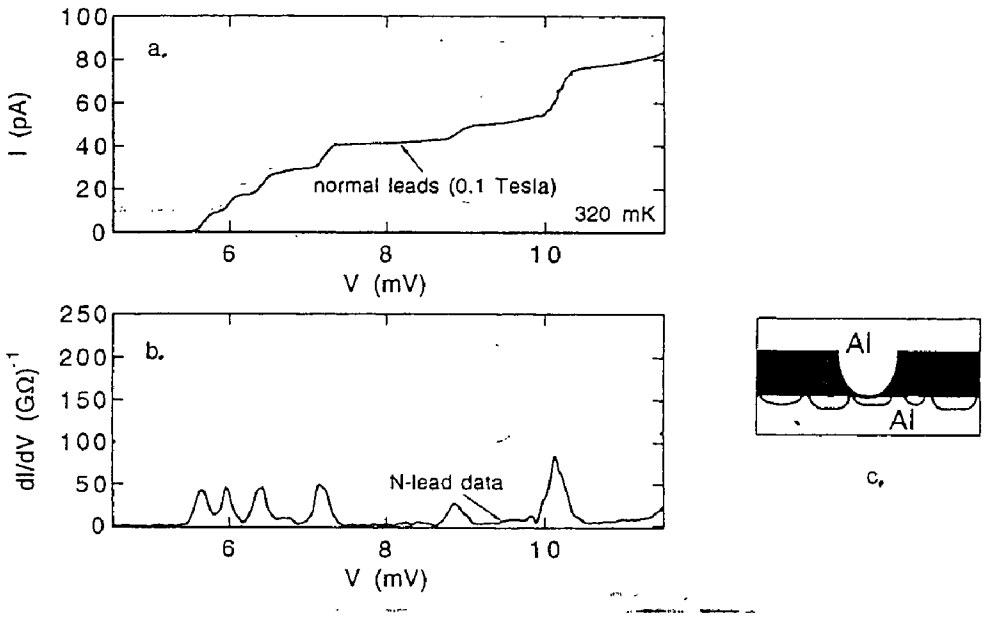


Figure 9.15. a. The I-V characteristic of the device at T=320 mK; the leads are forced normal by a magnetic field of ~100 mT. Note the Coulomb blockade threshold of ~6 mV. b. The derivative of the curve of figure a., highlighting the structure related to the confinement energies. c. Schematic diagram of the device, showing the small Al islands.

tunnel spectroscopy (see chapter 7 and 10) the electronic levels are extracted from the *I-V* (and d*I*/d*V*-*V*) curves. Figure 9.15a shows the structure of the device. It is made by etching a small hole (~ 5 nm) in a thin Si_3N_4 membrane. Next one Al electrode is evaporated, followed by an oxidation step: this forms the first tunnel barrier. Then on the side where the island has to be realised a very thin Al layer is evaporated of ~ 2 nm *on average*. The trick to form the island is now that the evaporation of very thin metal layers at room temperature commonly does not result in a closed and uniform coverage but instead a non-continuous, granular layer results showing the formation of small islands. Next this discontinuous layer is oxidised (yielding the second tunnel barrier) followed by evaporating the counter electrode. Note that the alignment of the etched hole and a metal island is just random, just as the shape and size of the island itself! To a lesser extent this also holds for the transmission of the barriers. The typical values found are ~ 10 nm diameter and a volume of ~ 100 nm^3, the last one yielding an average level spacing ΔE~ 0.7 meV; in addition a capacitance of ~ 15 aF (= $10*10^{-18}$ F) is estimated from the size, yielding a charging energy ~ 6 meV. Figure 9.15b and c show the results obtained at T=320 mK and a magnetic field of ~ 100 mT (to suppress the superconductivity of Al). Very clearly the Coulomb blockade threshold is seen at V=4-5 mV. Beyond this bias voltage a distinct step-like structure is seen in the current-voltage characteristic (= peaks in the differential conductance versus bias voltage). Note that the typical peak spacing between 5.5 and 11 mV equals ~ 0.5 meV, in reasonable accordance with the estimation based on the size of the island. It should be noted that quite a few more interesting results are presented in this paper, including effects due to the superconducting state.

The last experiment we want to discuss concerns a measurement of the magnetic susceptibility of Mg clusters, embedded in a frozen hexane environment (K. Kimuro et.al., Surf. Sc. 156, 883 (1985)). Note that Mg contributes an *even* number of electrons to the conduction band, so the system should be even. The magnetic moment is measured

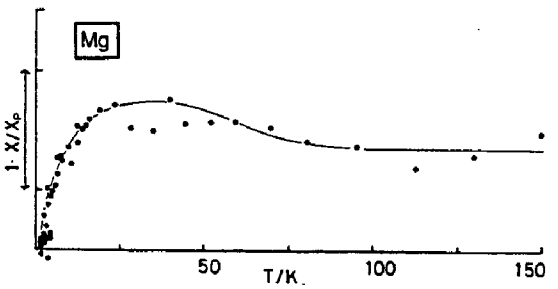


Figure 9.16. Magnetic susceptibility of Mg clusters in frozen hexane, measured using a Faraday balance. Mg particles of ~ 2nm diameter were used, with an average energy level spacing corresponding to ~ 250K. With each Mg atom contributing 2 electrons to the conduction band the total electron parity of each cluster is assumed to be even.

employing a so called Faraday balance, which measures the force resulting from the interaction between the applied field and the induced (spin-) moment. The typical size of the Mg particles is ~ 2 nm, resulting in a mean level spacing of ~ 20 meV (or ~ 250 K). Based on our preceding discussion (Fig. 9.12) we anticipate a suppression of the spin contribution to the magnetisation if $kT <\sim 0.5$ * the average level spacing, i.e. below ~ 120 K. Figure 9.16 shows the results of this experiment. Two features are evident: first a weak rise of X (with decreasing temperature) around ~ 40 K. Secondly, a clear reduction is seen to occur below ~ 50-60 K, which becomes even more pronounced if we "correct" for the hump around 40 K. The reduction of X is attributed to level repulsion and non-random level distribution, i.e. the effect we were interested in. The hump is not well understood and was attributed to some impurity effects.

9.6 Closing remarks

In this chapter we have discussed a number of aspects related to the electronic and structural properties of small conducting particles, called clusters. From the presented material it will be clear that the theory is rather well developed. Experimental verification of theoretical predictions have been largely hampered by the constraints posed by the sensitivities required to study clusters. One approach to overcome this difficulty has been to investigate large numbers of (nominally) similar clusters, but the unavoidable statistical variations between the members poses considerable limits to the interpretation of the results. The second approach, i.e. to study only single clusters, is only recently becoming into reach, and this only for a few quantities. Much experimental work is still to be done.

By no means full coverage of this broad subject is given in this chapter. Aspects like the recently discovered (and widely studied!) bucky balls and bucky tubes have not been discussed: scanning some recent issues of Nature, Science, Physics Today or the like will provide ample entrance into this area. Also the effect of confinement on the superconducting properties of small structures has not been included: here only few experiments exist. Some aspects are discussed in the reviews of Perenboom and Halperin.

General references to the subject

1. M.L. Cohen and W.D. Knight, Physics Today 43 (12), 42 (1990)
2. J.A.A.J. Perenboom et.al., Physics Reports 78, 173 (1981)
3. W.P. Halperin, Review of Modern Physics 58, 533 (1986)
4. W.A. de Heer, Review of Modern Physics 65, 611 (1993)
5. Various papers in: Science 271 (Febr. 16, 1996)

10. Scanning probe methods/techniques

Key words: scanning probing, STM, SFM, MFM, SIMS, ion beams, electron beams, Auger effect

10.1 Introduction

Mesoscopic physics concentrates on the properties of matter on a small sized scale. As we have seen in the preceding chapters this is commonly achieved by fabricating small structures, -such as rings (chapter 5), junctions (chapter 6 and 7), dots (chapter 7 and 8) or clusters (chapter 9) -, which are investigated e.g. via leads attached to the device. In a different approach one employs a particular lengthscale that is intrinsic to the phenomenon of interest, which allows to effectively "cut out" a section of this size from the macroscopic system. This is done e.g. in the case of weak localisation (chapter 2, with the phase coherence length as the scale of interest) or the quantum Hall effect (chapter 4, with the magnetic length as the typical length of confinement).
In the present chapter we will see how these methods can be complemented by techniques that allow the investigations of properties on an even more local scale. Some of these methods has been touched upon already before, but here we want to introduce these in a more systematic way.

The methods that we want to discuss go under the name of *local probing*. The degree of locality, i.e. the resolution in space, varies between methods, depending on the specific mechanism used, but it may reach values to *tens of pm*, i.e. well below the size of the smallest atoms! As in addition these techniques often allow the (nano-)probe to be directed to various positions on the surface of the object that is investigated, - indicated by *scanning* -, there is not too much imagination required to envision that the resulting *scanning local probing* is very powerful and versatile! This is even more so as some methods are also capable to be employed as tools to modify the surface on an atomic scale in a controlled way, in this way providing a means to perform *fabrication on an atomic level*.

We first (section 10.2) will provide a short introduction to various local probing methods based on the use of particle beams. Next we will discuss two recently emerged and by now very widely used methods, based on the use of local forces (section 10.3) and on local tunnel currents (section 10.4) respectively.

10.2 Particle beam scanning probe methods

The central issue in scanning probe methods and techniques is the intrinsic/generic capability of determining the properties of a surface on a very local scale. In this section

we will discuss a few methods that are based on the focused irradiation by particles of the surface under investigation, like electrons or ions. We will present a few methods employing particle injection, and see how this may provide detailed information on the local physical and chemical properties of the surface using various spectroscopic means.

In particle beam probing methods we can distinguish between two types of particles, namely electrons and atoms (or ions). In using electrons one commonly detects (in a spectrally sensitive way) the radiation that is emitted from the irradiated solid ("fluorescence"), and this provides direct information on the chemical nature of the material. In employing atoms or ions one commonly detects the particles that are knocked out from the surface by the incoming ions; by mass-spectroscopic means one can determine which atomic species are being emitted. For more details of this field the book of Fuchs (see end of this chapter) can be used.

10.2.a. Electron beams.

The basic idea here is to inject high energy electrons into the surface of the solid and see how the interaction of these electrons leads to energy transfer to the atoms constituting the matter under investigation. Figure 10.1 shows this process in a simple way. The incoming (high energy) electron penetrates the electron cloud surrounding the atom, and transfers its energy (partially) to an electron from one of the inner shells via inelastic scattering. This will remove this electron from the shell, leaving the atom in an excited state (i.e., with a hole at the place of the formerly bound inner electron). Next, this hole will be occupied by the transition of one of outer shell electrons to the inner shell hole. Now two processes may occur. Either the energy released at this transition is transferred to again another electron from one of the higher lying shells, leading to its emission as a so-called *Auger electron*, or it will result in the emission of a photon determined by the energy difference of the two bound states. For this second process the energy difference ΔE is *characteristic for the atom* hit by the incoming high energy electron, in this way allowing the determination of the particular atom from its emitted radiation. It may range up to ~100 keV for heavy atoms and so the resulting photons will be in the very short

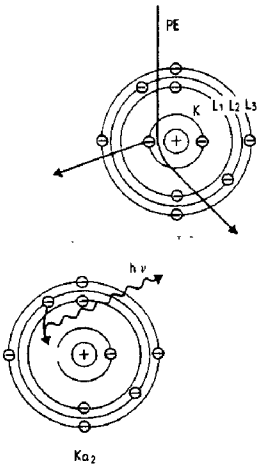


Figure 10.1. The process of the generation of an X-ray photon via the removal of an electron from the inner shell of an atom.

Figure 10.2. Microcomposition obtained from local X-ray analysis. The picture shows a superposition of a regular scanning electron micrograph and an X-ray image selectively measuring the local yttrium content (white "spots").

wavelength range of *X-rays*, following the relation $\lambda=hc/\Delta E$ (approximately an energy of 1 keV corresponds to a wavelength of 1 nm).
By scanning the incoming electron beam across the surface and measuring the emitted radiation in a wavelength-resolved way, the local composition of the surface can be determined, a method called *electron beam X-ray microanalysis*. Figure 10.2 shows a typical result obtained on a small area of barium titanate ceramic material. Here the X-ray detector was set to measure the characteristic X-rays emitted from Yttrium atoms. Note the lateral resolution of ~ 200 nm.
At first sight one may infer that the lateral resolution will be determined by the beamsize of the incoming electrons, which can be as small as 1 nm. This is, however, not so. The actual limit is governed by the penetration depth of the high energy electrons into the material, spreading these electrons also laterally over an area of the same order of magnitude. As these distances vary from ~10 nm (at ~ 1 keV or less) to ~1 μm (at ~ 100 keV), this will determine the lateral resolution.
Alternatively to employing the emitted X-ray, also the process which involves the emission of the (secondary) Auger electron can be used. In this case its energy allows information on its origin (i.e., from which atom it was emitted) to be obtained. We will not discuss this method here.
As a final note it should be indicated that employing these electron beam microanalysis methods will not affect (strongly) the properties of the solid under investigation. This follows from the smallness of electron mass relative to the atomic mass (and so the small transfer of its kinetic energy to the atom) and the relatively large binding energies of atoms in solids. In this respect the method is of a non-destructive nature.

10.2.b. Ion beams

The most important method employing ions as the incoming particles is called *Secondary Ion Mass Spectroscopy (SIMS)*. It employs the fact that when the solid surface is bombarded by a beam of primary ions, components of the surface are abraded by *sputtering*, and some of the removed particles are emitted as (secondary) ions into the free space above the surface. By collecting these secondary particles and investigating these by a mass selective method, the individual species removed from the surface can be analysed.

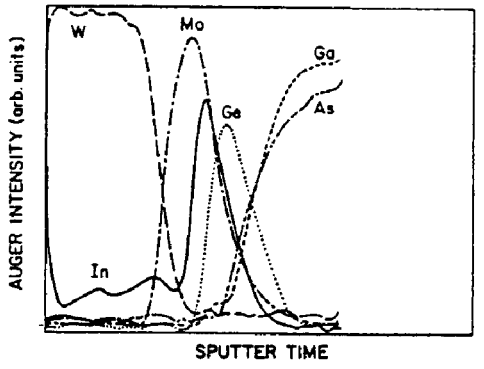


Figure 10.3. A SIMS depth profile for a GaAs substrate covered by a sequence of four metal layers: 10 nm Ge (at the GaAs), 2 nm In, 12 nm Mo and 20 nm W.

By scanning the (focused) beam of incoming ions across the surface, information on the local composition of the surface can be obtained.
From this short introduction we immediately can deduce the most important characteristics of SIMS. First, the sputtering action makes the method destructive in nature: the surface is directly modified by the abrasive action of the primary ions. Secondly, as ions only penetrate the solid very shallowly, the method is depth-selective to a few atomic layers (~ 1nm). Combining one and two immediately implies its use for depth-dependent investigations by the measurement of the secondary ions composition during the continuously further penetration into the material (in time) by the abrasive sputtering action. This provides a so-called *composition depth profile*. Figure 10.3 shows a typical depth profile of a semiconductor (GaAs), covered by a metalisation layer of Ge(10nm), In(2), Mo(12) and W(20). Evidently the time-axis is a measure of the depth reached during the sputter process. From the picture we nicely can see the transition from metal to metal and from the metal to the GaAs layer itself.
From the know metal layer thicknesses (as obtained by the metal evaporation parameters) this picture provides some typical numbers for the depth resolution that can be obtained. With the In and Ge peaks being comparable in width (= time) we may infer a resolution of ~10 nm. In view of the "fundamental" limit given by the ion penetration depth (a few nm), this may seem not so good. However, other contributions have to be taken into account, and these are much more restrictive. One important contribution is the effect of redeposition of the abraded material: a fraction of the material sputtered away from former layers is redeposited, and so this will contribute to the mass composition at lower layers of the material, in this way smearing the abruptness of any compositional interface transition. Secondly, any morphological variations along the surface (i.e. grains etc.) will result in different rates of sputtering and so fluctuations in the depth, leading to a spreading of the *average* composition seen by the *total* beam. Under favourable conditions effective depth resolutions of better than 10 nm can be achieved, comparing well with the example of figure 10.3.
The lateral resolution that can be obtained by SIMS largely depends on the diameter of the ion beam. This is governed by the design of the ion emitter and the ion-optics of the total beam generator: values of 50 nm or less have been attained for liquid metal ion sources.

10.3 Scanning force methods

Nature provides us with a wide variety of forces. Also in mesoscopic systems various types of force will be relevant. Whenever an atom is brought in close proximity to a second atom, interaction between the two will occur. This interaction will be either attractive or repulsive and so an (electrostatic) force between the two results. For the case of the persistent currents discussed in chapter 5 a finite magnetic moment results, and this will lead to a (small!!) magnetic force if a small magnet would be put close by.
In this section we will discuss scanning methods based on the use of local variations of force on a nanometer scale. Methods to investigate the submicrometer properties of (interfaces of) solids based on the use of these forces are denoted by *Scanning Force Microscopy* or *SFM*. Depending on the particular employed force names like AFM (atomic forces) or MFM (magnetic forces) are used. Given the small sizes involved, the forces commonly encountered are rather small: for the case of interatomic forces values as small as 10^{-10} N are quite normal, with the added characteristic of the very strong dependence on the distance between the two atoms. Needless to say that these very small values require special approaches to allow their proper measurement.

10.3a. *Basic properties*

In the world of atoms and molecules the electromagnetic interaction is the only relevant one of the four known fundamental forces in nature. Concentrating here on atomic interactions at short distances we want to gain understand of the modes of operation and the limitations of the (atomic) force microscope ((A)FM).
Basically the interatomic forces contain two contributions, one being relatively(!) long ranged and attractive in nature, while the second is considerably more short range and repulsive.
The *attractive* force is called the *Van der Waals* force and is a consequence of the dipole-dipole interaction between two atoms or molecules. To be somewhat more specific, the VdW force contains three types of contributions, one resulting from permanent dipoles, the second from dipole-induced dipole interactions and the last (commonly largest) arising from fluctuations in the atomic charge distributions (positive relative to negative). The VdW force behaves like r^{-7} for distance $\sim<10$ nm.
The *repulsive* force acts at a considerably shorter range, i.e. below a few tenths of nm. Two contributions can be mentioned. First the strong overlap of two electron clouds leads to incomplete screening of the nuclear charges, resulting in a strong coulombic (repulsive) force between the two atoms. Second, the Pauli exclusion principle will try to avoid electrons to occupy the same volume in space, in this way also giving rise to a repulsive action.
Figure 10.4 shows a typical curve of the total interaction energy (with the resulting force being the gradient of this energy) in dependence of distance r, showing the characteristic Lennard-Jones shape. It will be evident that the long range nature of the Van der Waals forces will not favour atomic resolution to be achieved by the scanning force microscope: the long tails of the forces (~1 nm or more) of the many atoms on the surface interacting with those in the scanning point-shaped probe (or tip) will simply average out all lateral atomic-sized fluctuations along the surface of interest with a typical scale of ~0.1 nm.

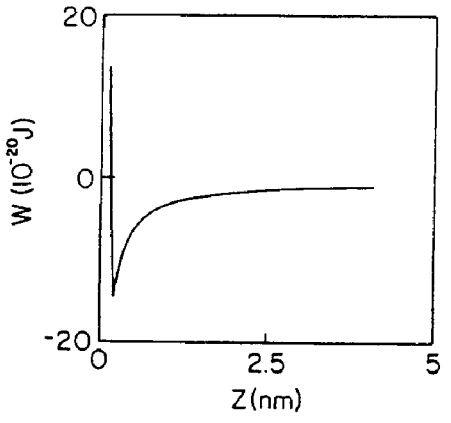


Figure 10.4. Schematic representation of the intermolecular Lennard-Jones type of potential U(r), for the case of two atoms or molecules.

Only the short range repulsive interaction allows the full, atomic resolution to be obtained from the SFM. This implies that, in order to attain atomic resolution, an (atomically) sharp tip should be brought in very close proximity to the actual surface, i.e. to operate the AFM in the *contacting mode*.

One more aspect of the force-distance relation has to discussed. If the distance is reduced more and more, the resulting rapid increase of the interaction will lead to *elastic* modifications (=reversible) and even *plastic deformations* (=irreversible) of the surface (and the tip as well!), and may ultimately lead to a direct contact: the two stick together. Forces of $\sim 1*10^{-7}$ N seem to be enough to achieve this last condition. Evidently this last occurrence will result in surface damage. Rephrasing this consequence more positively however, means that this working condition can also be employed for the purpose of *nanostructuring* and some words will be devoted to that in section 10.5.

10.3b. *Experimental aspects*

The basic operation of the Scanning Force Microscope shows considerable resemblance to that of the STM (see section 10.4). Figure 10.5 shows the typical diagram for its operation. The *sharp tip* acted upon by the applied force is mounted at the end of a *deflectable lever*. Piezo-electric transducers, which provide a well controlled elongation (and so displacement) at the application of a given voltage, are used for vertical and lateral displacement. The vertical displacement of the lever arm is sensed and used as the feedback information for the *z*-piezo, which keeps the force between sample and tip constant. In this way the voltage at the *z*-piezo directly reflects the size of the local (force-)corrugation ("hills and valleys"). The lateral displacement is controlled via the *x*- and *y*-piezos, allowing the scanning of the surface.

Evidently the tip and cantilever are crucial for the proper operation. Basically the tip shape largely determines the lateral resolution that can be attained. In addition the (elastic) properties of the cantilever, including the detection of its vertical displacement, are central for the vertical resolution.

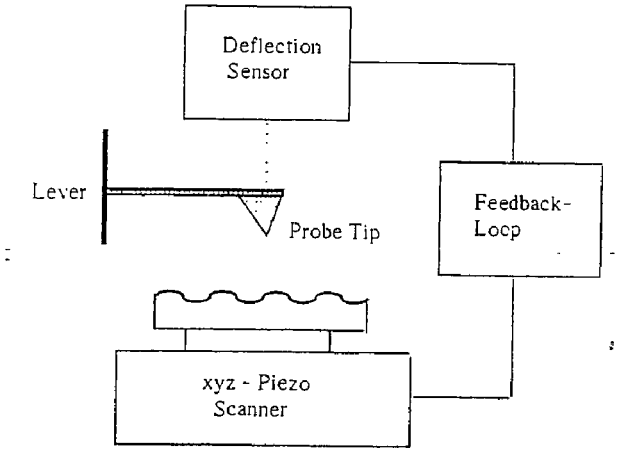


Figure 10.5. Basic diagram of a scanning force microscope. The sharp probing tip is attached to cantilever-type spring. The deflection resulting from the measured force is sensed, and fed back to the z-piezo, maintaining a constant tip-sample distance (or force). The x- and y-piezo allow the lateral scanning of the sample surface.

Many different schemes are used for this vertical detection, based on optical, capacitive and electron tunneling ((S)TM "piggy-backed" onto the SFM!) methods. Recently the use of piezo-resitive (i.e. showing a strain-dependent resistance) cantilevers has emerged rapidly. Figure 10.6 shows the basic aspects of this type of detection. The resistive cantilever (made of doped Si) has two legs (Fig 10.6a), which allows its resistance to be measured. Given the strain-dependent nature of the material, the resistance directly reflects

a. b.

Figure 10.6. Piezo-resistive cantilever employed for scanning force microscopy. a. The two-legged cantilever beam, with the two electrical contacts for the measurement of its resistance at the right. b. The tip, located at the endpoint of the cantilever: total height ~ 3 μm.

the "vertical" deflection of the cantilever, and so of the force acting on the tip (provided the elastic constants of the cantilever beam are known). Note the (barely distinguishable) tip at the end of the cantilever in figure 10.6a. Figure 10.6b provides a SEM picture at a much larger scale of the actual tip; with ~ 3 μm total height a tip radius of ~ 20 nm can be achieved. Typical force constants of ~ 10^{-11} N/nm are (commercially) available, and combined with a resolution of one part in a million for the resistance measurement it allows the determination of forces well below 10 pN. All these numbers imply that sub-nm vertical resolution is possible.

10.3.c. *Experimental results*

In this subsection we will present three characteristic examples of measurements performed by scanning force microscopes. These will range from the atomic arrangement of a crystal, the magnetic arrangement on a computer hard disc, to the imaging of a large organic molecule.

Figure 10.7 shows a crystalline surface which is rather well known in the field of SFM and STM, namely graphite. The spacing between atoms is approximately 0.24 nm. With a load force of 10 nN a vertical corrugation of ~ 0.02 nm is measured.

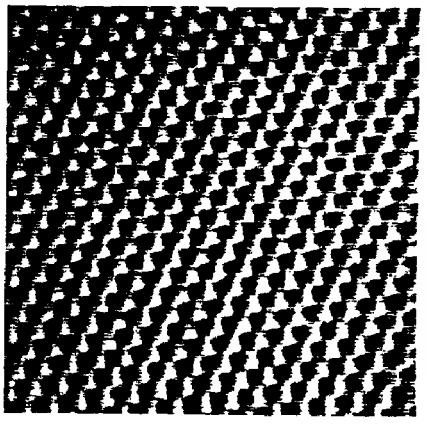


*Figure 10.7. Scanning force image of a graphite surface, showing atomic resolution. The (vertical) corrugation is 0.02 nm, and the applied loading force is 10^{-8} N. Total area is 4.0*5.5 nm^2.*

The second example concentrates on the measurement of magnetic forces. Figure 10.8 shows an image obtained from the magnetic ordering similar to that found at the surface of a (computer) hard disc. The magnetic recording track has a width of ~10 μm, with magnetic transitions every 5 μm. The image is obtained with a Ni tip. From this image we can deduce a resolution of at least 100-200 nm. Note the "force roughness" or "force noise" across the surface. This is not due to topological corrugations but indeed represents the noise occurring in magnetic image itself, demonstrating the limits for the signal-to-noise ratio that can be attained in such systems.

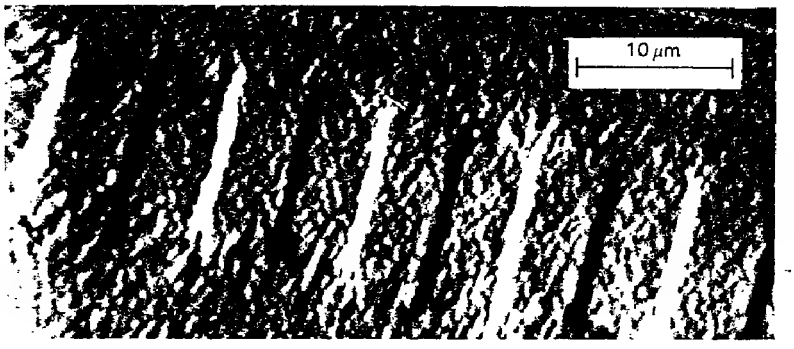


Figure 10.8. Magnetic force microscope image obtained with a Ni tip on a magnetic record track on a CoPtCr medium. The tip was polarised in the direction perpendicular to the sample surface.

From various experiments performed on magnetic systems with very abrupt local changes of the magnetic polarisation (e.g. at the walls between neighbouring magnetic domains) an experimental lateral resolution of ~ 10-15 nm has been achieved. From rather simple mode descriptions it is clear that this is mostly determined by the tip radius (~10 nm) and the tip-sample distance.
The third example originates from a completely different area and again relies on atomic interaction. Figure 10.9 shows an AFM image obtained from a DNA string deposited onto a metal surface. It is taken in the non-contacting mode, i.e. at a relatively large distance from the sample. The apparent width of the string deduced from the image is

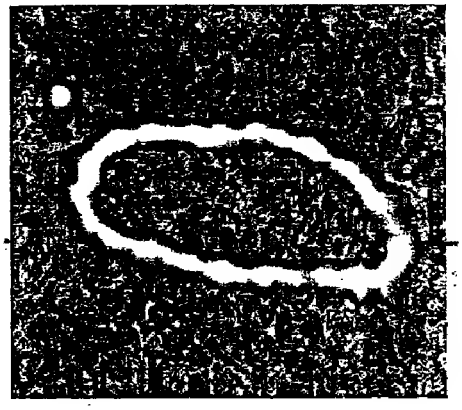


*Figure 10.9 Atomic force microscopic image obtained from a DNA string onto a metal substrate. The image is taken in the non-contacting mode. Total field of view is 1*1 μm^2.*

approximately 30-50 nm. Note that this is much larger than the known diameter of a bare DNA string, which is ~ 2nm. In the present image the measured width is determined by the relatively large distance between tip and molecule (due to the non-contacting mode of operation) and the tendency of the molecule to entangle once put into "free space".

Before closing this section on force microscopy some final remarks may be in place. The absence of nearly any preconditions related to surface properties evidently underlines the broad applicability of the SFM. This has been materialised over the last ten years as it has been used in fields like chemistry and electrochemistry (like catalyses), biology and medical sciences, in applied sciences like electronics, magnetics (like recording) and physical chemistry (e.g. polymers), with in addition a variety of applications in pure science.

10.4 Scanning tunneling methods: microscopy and spectroscopy

In this section we will concentrate on the transport of electrons between two conductors that are very weakly coupled. This separation could be realised by a narrow gap of vacuum or an insulator. The resulting barrier classically will prevent current to flow between the two pieces of material. However, as we have seen before (see chapter 1 and 9) quantummechanically this is still allowed by the process called tunneling. It is our aim to see what type of information can be gained from the tunneling into a local position on a surface and how this can be accomplished experimentally.
In section 7.2 we have discussed tunneling in the context of the controlled exchange of single electrons between leads and a small island. In that particular case tunneling was only employed as a way to "isolate" the island from the leads to such a degree that the (wavefunction of the) electrons can be assumed to localised, i.e. the electron either sit on the island or in one of the leads. In the present section we are more interested in which effects determine the current flow in the tunneling process.

10.4.a. *Theory of tunneling*

Figure 10.10 shows the energy diagram (or landscape) resulting from the formation of a tunneling system by bringing two conductors (1 and 2) in close proximity. If the two are put together in such a way that weak coupling arises, electrons can be exchanged between the two and so in equilibrium this results in the alignment of the electrochemical potentials of the electron reservoirs. The (different) workfunctions Φ associated with the two conductors determine the barrier height and shape. Note that 1 and 2 can be a normal metal, semiconductor, or a superconductor.
To obtain a quantitative description that would allow the calculation of the current in dependence of voltage, temperature, tunnel barrier properties, electronic properties of the conductors etc., basically two approaches can be taken. One way is to take an electron wave incident on one side of the barrier and to calculate its transmission by matching the wavefunctions at the two boundaries at each side of the barrier. The transmitted waveamplitude directly provides the current flowing via this tunneling wave. It is called

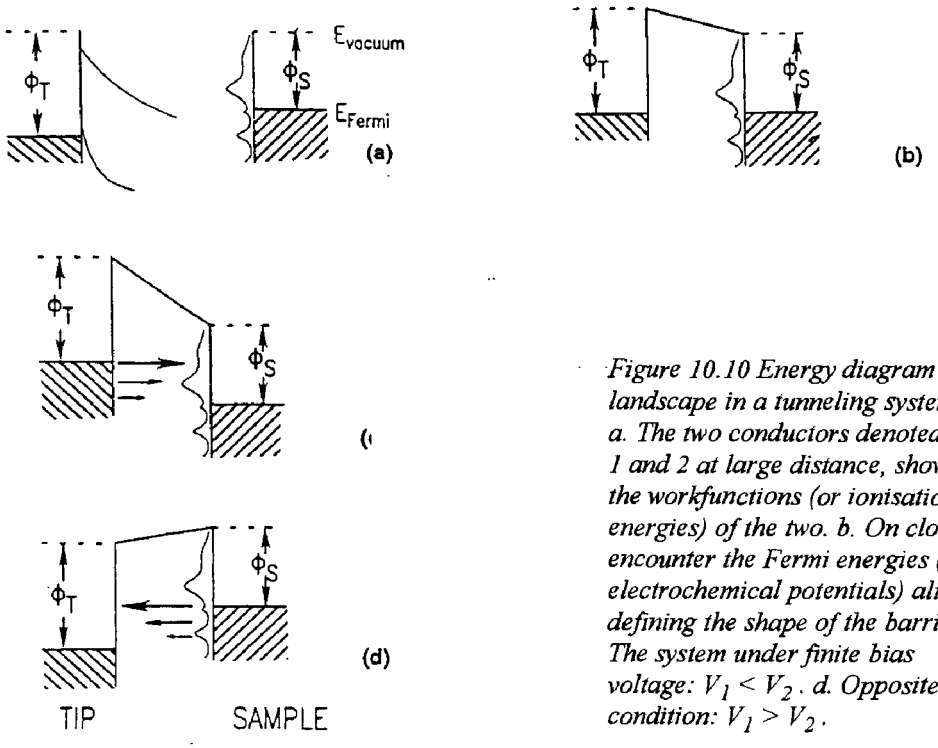


Figure 10.10 Energy diagram or landscape in a tunneling system. a. The two conductors denoted 1 and 2 at large distance, showing the workfunctions (or ionisation energies) of the two. b. On close encounter the Fermi energies (or electrochemical potentials) align, defining the shape of the barrier. c. The system under finite bias voltage: $V_1 < V_2$. d. Opposite condition: $V_1 > V_2$.

the wave-matching method. In the second approach, called the transfer or tunneling Hamiltonian method, the barrier is explicitly assumed to be of small transparency, allowing the use of perturbation theory in the problem. Now the overlap matrix element of any two wavefunctions at different sides of the barrier is calculated, from which the current (transferred via that pair of states) can be evaluated.
Employing the second approach, one describes the system by two Hamiltonians for the unperturbed eigenstates of the two conductors, and a perturbation term H_t associated with the weak tunneling coupling. Now we can apply the famous Fermi's Golden Rule to evaluate the probability for the transition (i.e. the transition or tunneling rate) from one particular initial state i_1 in reservoir 1 at one side of the barrier to a state j_2 in reservoir 2 at the other side. This reads

$$\Gamma_{i_1 \to j_2} = \frac{1}{h} |< i_1 | H_t | j_2 >^2 \delta(E_{i_1} - E_{j_2}) \qquad (10.1)$$

where the matrix element describes the overlap between the wavefunctions of the two states and the δ-function guarantees that the process is elastic (i.e., no loss of energy *during* the actual tunneling process itself). This energy conservation basically says that the

two states have to be aligned to allow the transfer of the particle/wave. Evidently, to obtain the actual rate i_1->j_2 one has to multiply by the degree of occupation of the 1-state, f_i, and that of vacancy of the 2-state, (1-f_j). This yields

$$\Gamma_{i_1 \to j_2} = \frac{1}{h} |< i_1 | H_t | j_2 >^2 f_{i_1}(1 - f_{j_2}) \delta(E_{i_1} - E_{j_2}) \qquad (10.2)$$

To obtain the total rate of particles flowing between the two reservoirs 1 and 2 we of course have to sum over all states i_1 and j_2. To do so we assume that the density of states ρ_1 and ρ_2 at sides 1 and 2 is large. This allows us to rewrite the sum over all states as an integral.

Applying a bias voltage V between the two conductors shifts the Fermi levels by eV. Limiting ourselves to rather small energy shifts implies that only states close to the Fermi energy are of importance. As, in addition, the large DOS implies a strong mixing between states, it is very likely that the character of all the states is rather similar. This allows the matrix elements to be replaced by the product of an average transmission probability $T_{12}=|t_{12}|^2$ (with t_{12} being the transmission coefficient through the barrier). The result for the integrated current from 1 to 2 now can be written as (see eq. (7.8))

$$I_{1\to 2} \equiv e\Gamma_{1\to 2} = \frac{e}{h} \int_{-\infty}^{\infty} T_{12} \rho_1(E - eV) f(E - eV) * \rho_2(E) \{1 - f(E)\} dE \qquad (10.3a)$$

For the current flowing in the reverse direction from 2 to 1 we obtain in the same way

$$I_{2\to 1} \equiv e\Gamma_{2\to 1} = \frac{e}{h} \int_{-\infty}^{\infty} T_{12} \rho_2(E) f(E) * \rho_1(E - eV) \{1 - f(E - eV)\} dE \qquad (10.3b)$$

If the bias is taken such that $eV<0$ (see figure 10.1c) at zero temperature this reverse current will drop to zero as *all occupied* states in 2 that could contribute to the current towards 1 are faced with occupied states in 1. So, for the total current only the contribution from 1 to 2 needs to be taken, integrated over the bias energy interval eV, which yields

$$I_{12} \equiv I_{1\to 2} - I_{2\to 1} = I_{1\to 2} = T_{12} \frac{e}{h} \int_{E_F}^{E_F - eV} \rho_1(E - eV) \rho_2(E) dE \qquad (10.4)$$

In the limit of very small bias voltage we may assume that the DOS at both sides will be approximately constant within this energy range and so we find for the *linear regime*

$$I_{12} \equiv T \frac{e}{h} \rho_1(E_F) \rho_2(E_F) \Delta(E) = T \frac{e^2}{h} \rho_1 \rho_2 V = GV \qquad (10.5)$$

with G denoting the linear conductance of the tunnel junction.

If the DOS does depend on energy the (incremental) current will become bias dependent. For this more general case from eq. (10.4) we simply can evaluate the differential conductance at a given bias voltage

$$\frac{dI}{dV} = T\frac{e^2}{h}\rho_1(E-eV)\rho_2(E) \qquad (10.6)$$

The transmission T through the barrier can be evaluated approximately, assuming that the states can be represented by plane waves and taking a rectangularly shaped barrier of width d_b and height E_b (relative to the Fermi energy). From textbook quantummechanics we know that the wave penetrating the barrier will show an exponential decrease, with a decaylength given by $l_b = 2\pi/\kappa = h/\sqrt{2mE_b}$. The overlap between the plane waves at both sides of the barrier now approximately equals $T_b \approx t_{12}^2 \approx exp(-2\kappa d_b)$, i.e. it goes with the square of the transmission coefficient. To gain some feeling for typical numbers, we take the barrier height E_b approximately equal to the workfunction (~ 5 eV for a metal versus vacuum); with a barrier of the size of just one atomic layer (~ 0.2 nm) we find T_b~0.01, i.e. << 1. Note that this small value resulting from the very thin barrier should be treated as a very rough estimate, as it is by no means evident that the "macroscopic average" workfunction holds for such sub-nm situations.

The important conclusion that can be drawn from the preceding discussion is that the total tunnel current depends directly on the density of states at the two sides of the barrier. More particularly, this implies that if the DOS at one side, e.g. ρ_1, is known (or at least is constant), eq. (10.6) immediately shows that the differential conductance is a *direct measure for the local DOS* $\rho_2(E)$, including its *dependence on energy*. This forms the crucial aspect defining the intrinsic capabilities of the technique of tunnel spectroscopy. If one of the electrodes is made needle-shaped and allowed to be positioned at different places across the surface of the second electrode, the resulting system is called a *Scanning Tunneling Microscope/Spectroscope* or *STM/S*.

10.4.b. *Experimental aspects*

From the preceding section it is clear that for tunneling to occur the barrier between the two electrodes has to be in the range of atomic sizes, because of the exponential decrease of the overlap between the electron states at the two sides. However, this only is somehow a minimum requirement: a stable measurement of the current (say to better than 1 %) at a given voltage dictates the exponent to be stable to this value, and so the width of the barrier (or the distance between the two conductors) should be stable within ~< 1 pm (or 0.001 nm)! From now on we assume that one of the electrodes is configured as a very sharp needle, or tip, while the second electrode is taken to be "flat" (apart from atomic scale corrugations) and is denoted the substrate. For the ease of reference we will denote the direction perpendicular to the substrate as vertical (or z) and along it as lateral (or x,y). In order to explore the limits

further we take the tip to be of the size of a single atom. If this is so it is not difficult to see that in the lateral direction a resolution of the order of the tip size should be attainable. Evidently this implies that, for a stable current, also in the lateral direction the tip should not be allowed to move (unintentionally!) by more than ~ 1 pm. So we have to conclude that tip-to-substrate displacements should be controllable to much less than interatomic distances. If we succeed to do so, the road is open to investigate electronic properties of surfaces on a sub-atomic length scale, in particular if also the tip can scan the surface of the substrate.

Figure 10.11 shows schematically the setup for such a *Scanning Tunneling Microscope* (compare to the rather similar diagram of the AFM, figure 10.5). In figure 10.11a the tip is seen to be mounted on a carrier that can be displaced by applying voltages to the three attached piezo elements *x,y,z*. The two insets show the configuration of tip to substrate at increased magnification. Figure 10.11b shows the electronic circuit employed for the STM. The translation or scanning in the lateral direction is driven by the computer. While scanning the feedback can be operated in two modes. If the feedback loop is closed it will effectively make the *z*-piezo to move such that the current sensed by the tip is kept constant: this implies that it keeps a *constant distance relative to the local corrugation* (e.g. atomic "hills and valleys"). The feedback voltage operating on the vertical piezo is a direct measure of this corrugation. This mode is referred to as the *constant current* mode. In the second way of operation the feedback loop is opened. As the tip now hovers above the surface at a constant *average* height the current entering into the tip directly reflects the local structure of the surface. This way of operation is denoted as the *constant height* mode. It should be kept in mind that of course variations in the local density-of-states will affect the current(-variations) measured during scanning.

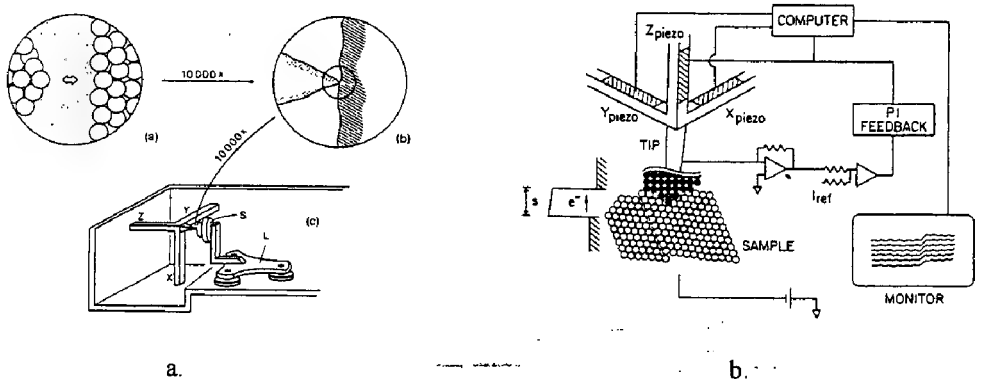


Figure 10.11. Schematic diagram of a scanning tunneling microscope. a. Shows the tip mounted onto a piezo driven construction, allowing displacement in three orthogonal directions via potentials to the piezo elements. b. Shows the electronic feedback circuit that allows the distance from tip to substrate to be kept constant, e.g. during a lateral (x,y) sweep along the surface.

Both methods presented above perform a measurement of the current at constant tip-substrate bias voltage V, i.e. including only the contribution of states within a constant energy window $\Delta E=eV$. From eqs. (10.4) and (10.6) we can see that in order to obtain information on the energy dependence of the density of states $\rho(E)$ the voltage should be swept and the resulting I-V curve will contain the required information. More specifically, if we assume that the density of states of the tip (say $\rho_2(E)$) can be taken to be approximately independent of energy (which is the case if it is a normal metal, with the Fermi energy somewhere in the middle of an energy band), the derivative of the current (10.6) with respect to the voltage immediately yields

$$\frac{dI}{dV} \cong T\frac{e^2}{h}\rho_1(E)\rho_2(E_F) = const.\,\rho_1(E_F - eV) \qquad (10.7)$$

The assumption of a constant DOS of the (metallic) tip holds quite well if the substrate is a semiconductor or a superconductor, as the range of energies of interest is comparable to the gap, i.e. ~ 1 eV for semiconductors and ~ 1 meV for superconductors. Note that if the density of states of the tip can not be taken as constant one will find a convolution of tip and substrate states, a case which will make the interpretation of results more complicated. Given the energy resolving capability resulting from this approach it is called *tunneling spectroscopy*.

Before proceeding to the discussion of some experiments one word needs to be devoted to the conditions of the surface. Evidently, to investigate the properties of a surface it is essential that the surface is well-characterised, i.e. it should be *atomically clean* and *free of contaminations*. As this may seem to be completely obvious, it is exactly this aspect which is one of the major problems encountered in STM and STS (= ST Spectroscopy). To obtain reliable experimental data this often implies that the STM needs to operated in a ultra high vacuum (UHV) environment to prevent the formation of adsorbates and oxide films. Often this requires that the surface to be investigated has to formed in UHV, e.g. by cleaving the material in the vacuum chamber.
Evidently the same thing holds for the tip: also this needs to be well prepared and clean. In addition the required atomic resolution requires that the shape of the tip needs to be point-like on an atomic scale. This guarantees a well-defined atomic state (or orbital) to be available to tunnel from. The appropriate preparation of the tip (e.g. by electrochemical etching) is an important step for successful STM experiments.

As a final remark we should stress the versatility of the STM method by noting that it can also be used to study dynamical phenomena occurring at the surface. As scanning can be performed at rather high rates, the change of local properties can be investigated during time. This allows e.g. the study of growth mechanisms in real time.

10.4.c. *Experimental results*
In this subsection we want to present a few characteristic results obtained at both low voltage in the linear regime, as well as at higher bias to perform spectroscopy. Metals, semiconductors and superconductors will be discussed in some detail, while one example of a biological nature will be mentioned.

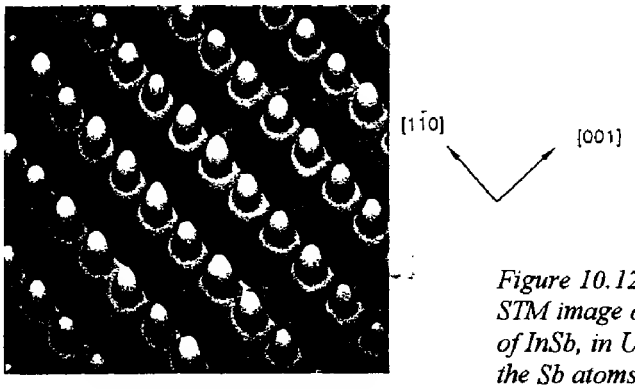


Figure 10.12. Constant current STM image of the (110) surface of InSb, in UHV. The light dots are the Sb atoms.

Figure 10.12 shows a nice STM image of a clean semiconductor surface; the topographic characteristics are enhanced (by computer) by representing the data both in grey-scale as well as in a pseudo-3D perspective manner. Taken in the constant current mode it shows

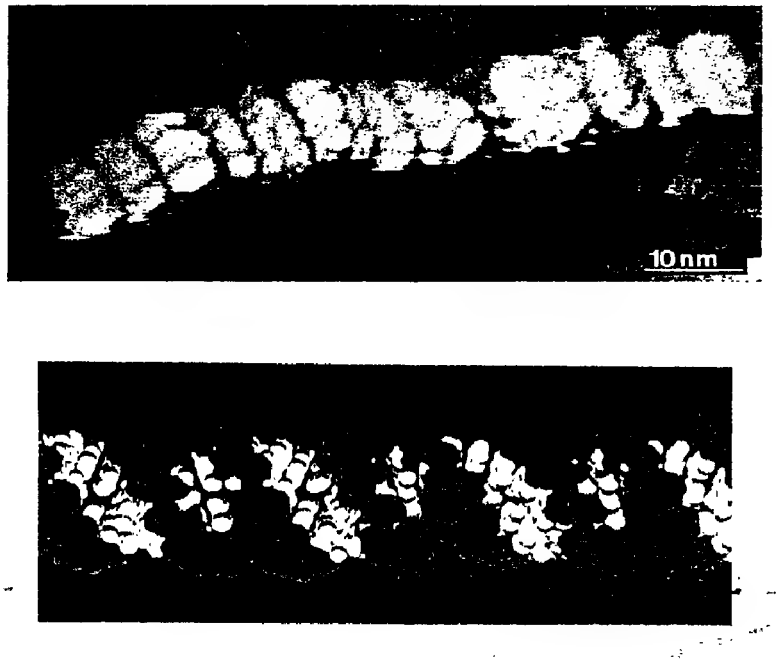


Figure 10.13. Part of a DNA molecule deposited onto a PtC film pre-treated by Mg salts. a. The actual image, with each section being a half of the DNA helical turn; b. For comparison a model of the DNA macromolecule is shown, with a total length of 12 nm.

rows of atoms at the (110) surface of InSb, taken in UHV (Y. Liang et.al., J. Vac. Sc. Techn. **B9**, 730 (1991)). Taken at a current level of 0.1 nA, with electrons flowing out of the substrate (i.e. the substrate at a -0.2V versus the tip), it shows the Sb atoms as light dots. Note the details in the grey-scale image, indicating a lateral resolution which is well below the interatomic distance.
As a second example figure 10.13 shows a bare DNA molecule adsorbed to a PtC surface (M. Amrein et.al., Science **243**, 1708 (1989)). Figure 10.13a clearly displays sections, each section being one half of the helical turn characteristic of the DNA molecule (see figure 10.13b; not on the same scale as 10.13a!). This image is obtained with the PtC surface pre-treated by Mg salts, and held under slightly wet conditions. Without discussing this any further this implies that STM also can be employed to study a variety of (electro-) chemical processes and mechanisms under aquatic environmental conditions.

To further illustrate the capability of STM to study the local DOS figure 10.14 shows results obtained on a GaAs/AlGaAs multi-layer heterostructure (see chapter 1 and 3 for more details). In such a system the edges of the conduction band (CB) and the valence band (VB) show characteristic offsets due to the difference in electron (or hole) affinity of the two materials. This step is known to be ~ 250 meV for the conduction band (with the AlGaAs being higher than the GaAs), and ~ -200 meV for the valence band (with AlGaAs lower than GaAs; the total difference in bandgap follows as ~ 450 meV). The top of the figure shows the layer structure, grown by MBE starting from the n-doped GaAs buffer followed by sequentially adding ~100 nm thick layers of n- or p-type AlGaAs and GaAs. The sample was prepared by cleaving the grown layerpackage perpendicularly to the growth direction and passivating the surface by dipping into ammonium sulphide. The

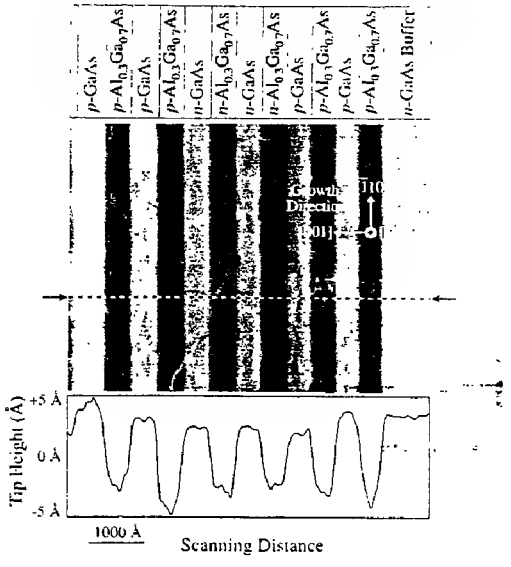


*Figure 10.14. A 700*500 nm^2 STM image of an $Al_{0.3}Ga_{0.7}As/GaAs$ multilayer heterostructure, obtained with a sample bias of -2.3 V and at a constant current of 0.3 nA. The top part shows the layerstructure; the lower part shows the tip distance while scanning the cleaved surface The main part of the figure shows the stripes representing the difference in valence band offsets of AlGaAs (dark) relative to GaAs.*

lower two parts of the figure show the result from the STM experiment done in UHV conditions (S. Gwo et.al., Phys. Rev. Lett. **71**, 1883 (1993)), taken at 0.3 nA and with the sample set at -2.3 V relative to the tip. This implies that the Fermi level of the tip is well inside the CBs, and so electrons will be transferred from the semiconductors into the tip. With the AlGaAs VB lying lower than that of GaAs the total energy interval of filled VB states of GaAs will be larger than of AlGaAs. If the tip would be held at a constant distance relative to the semiconductor interface this would imply a smaller current when scanning the AlGaAs as compared to the GaAs. Or, if operating in the constant current mode, the tip will move closer to the AlGaAs than to the GaAs. This is exactly what is seen in the lowest part of figure 10: while crossing the AlGaAs layers the tip hovers across the interface ~ 7 nm closer than for the GaAs sections.

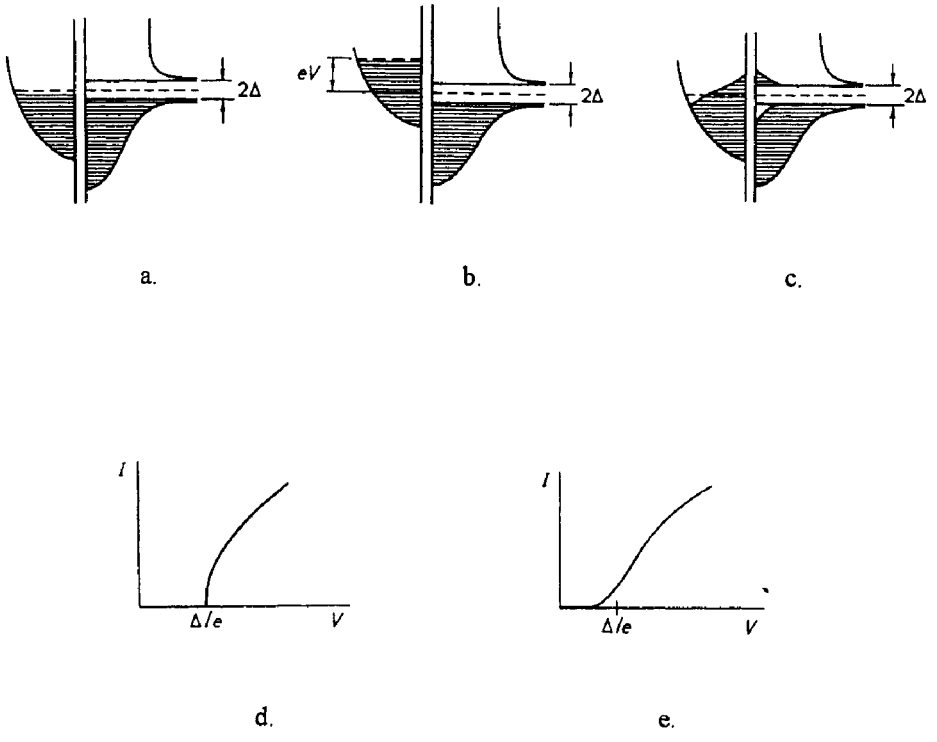


Figure 10.15. Density-of-states versus energy diagrams and I-V curves for tunneling between a normal conductor and a superconductor. a. The diagram for zero bias voltage (i.e. Fermi energies are aligned); normal metal left, superconductor (with gap 2Δ, right); note the enhanced DOS in the superconductor just above and below the gap. b. Diagram at bias V>Δ/e. c. At T>0 K the Fermi-Dirac distribution broadens the transition from empty to occupied states. d. The I-V curve at T=0; note the abrupt rise of the current at the gap. e. The effect of T>0 is a more gradual increase of the current at the gap.

In chapter 6 we have introduced superconductivity. One of the most characteristic features of a metal in the superconducting state is the formation of an energy gap around the Fermi energy, resulting from the condensation of "normal" electrons into pairs of electrons, called Cooper pairs. The energy gap denotes the absence of single particle states. Evidently this should show up in a tunneling experiment as a reduction of the current for voltages corresponding to energies within the gap (typically ~ 1mV).
Figure 10.15 shows the DOS-diagrams from which we can deduce the most characteristic features of the *I-V*-curve. Figure 10.15a and b show the $\rho(E)$ diagrams for the normal metal N (left) weakly connected to the superconductor S (right) via a tunnel barrier. With the bias $|V|<\Delta/e$ no empty states are available in S ($\rho_S(|E|<\Delta)=0$) , and so no current will flow (see eq. (10.5)), as depicted in figure 10.15d. At the gap the discontinuous increase of ρ_S to a value far beyond the normal state value leads to a steep rise of the current, followed by a more gradual increase when the DOS in S returns to that of the normal metal. In figure 10.15c the effect of the temperature is seen to broaden all features via thermal smearing. The resulting *I-V*-curve for $T>0$ is shown in figure 10.15e, with a much more smooth rise of the current near the gapvoltage. For the differential conductance $dI/dV(V)$ given by eq. (10.6), reflecting the DOS of the superconductor, we immediately infer that this conductance will be zero within the gap (1), followed by a peak-shaped rise close to the gap voltage (2) and a reduction to a voltage independent value beyond the gap (3). Figure 10.16 shows an experimental result of this nature. It is obtained on $NbSe_2$, a superconductor with a critical temperature of 7.2 K and with excellent surface quality properties.
An attractive feature of this $NbSe_2$ system is that it is a type II superconductors This implies that the application of a magnetic field affects superconductivity in a non-uniform manner, giving rise to a regular lattice of "normal" areas, called vortices, embedded in a superconductive environment. Increasing the magnetic field will increase the vortex

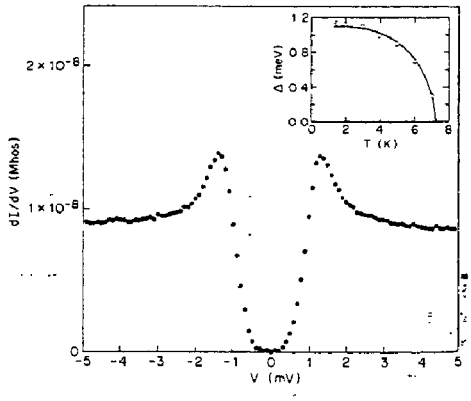
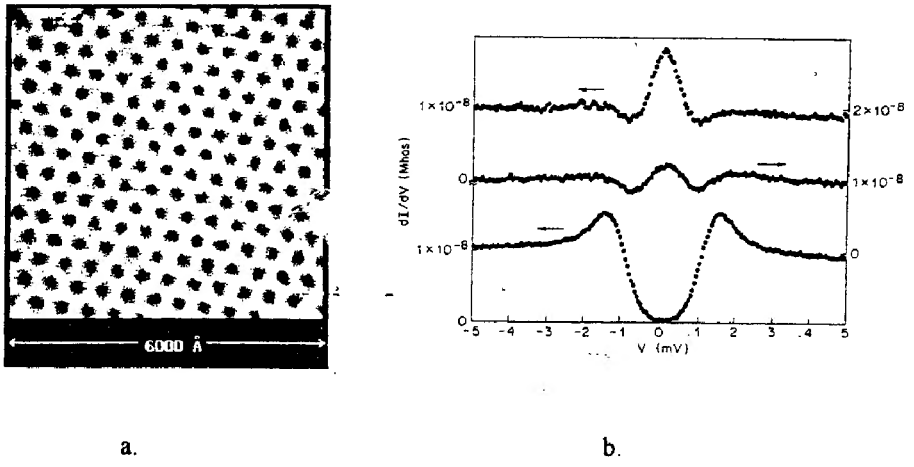


Figure 10.16. The dI/dV-V characteristic for $NbSe_2$. Note the good agreement with the derivative of the I-V curve shown in figure 10.15e, with numbers 1, 2 and 3 corresponding to those in that figure. The inset shows the gap versus the temperature, with T_c ~7 K.



a. b.

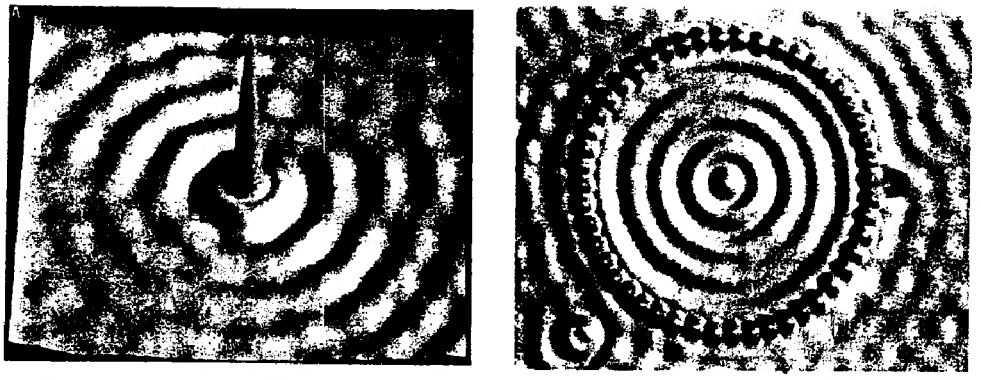
Figure 10.17. STM into a vortex lattice induced in the type II superconductor $NbSn_2$. a. The Abrikosov vortex lattice formed at a field of B=1T and at T=1.8K; white indicates the superconductor. The vortex rows are ~40 nm apart and each vortex is ~15 nm in diameter. b. The differential conductance dI/dV-V at three positions relative to a single vortex: the topmost curve is in the centre of a vortex, the mid curve at 7.5 nm from the centre (i.e., "at" the vortex edge), and the lowest curve inside the superconductor (see figure 10.7); three curves are offset for clarity.

density, ultimately leading to a complete suppression of superconductivity once all vortices overlap. Figure 10.17a shows an STM image of such a so called Abrikosov flux or vortex lattice in $NbSn_2$ in a magnetic field of ~ 1T and at T=1.8K. Note the nice regular triangular pattern of the vortices. The vortex rows are 40 nm apart, with the vortex diameter being ~15 nm.

Figure 10.17b shows the spectroscopic *dI/dV-V* curves taken at three different positions relative to a single vortex. Just outside of the vortex, i.e. in the area where the material is superconductive, a curve very similar to that of figure 10.16 is found (lowest curve), indicating a strong suppression of the DOS around the Fermi energy. In the mid curve taken near the edge of vortex, a significant change is evident: now the DOS is non-zero for all energies, and a small feature is seen near zero energy. The topmost curve is taken at the heart of the vortex, where the gap is strongly suppressed due to the penetrating magnetic field. Now a large *peak*-shaped feature is seen at the position of the *dip* in the superconductor. It took some time before the origin of the peak became clear. One can model the system as a normal vortex core surrounded by a superconductor. From what we have seen in chapter 6 this implies that electrons residing inside the core will experience Andreev reflection if approaching the superconducting wall around. As the reflected AR holes will be reconverted into electrons at "the other side" of the vortex (etc...), this will result in coherent multiple AR and so to the formation of Andreev eigenstates, much in the

same way as we have seen before. The features seen in the differential conductance curves indicate the *discrete DOS* resulting from these eigenstates.
Just as a short remark: note that this experiment does not require the STM to have atomic resolution in the lateral directions, as the vortices are large. However in the vertical direction the requirements are as usual, to ascertain appropriate current stability for the spectroscopy.

The last experimental result we want to discuss is at the heart of every physicist. It images in a direct way the spatial distribution of the quantum probability of electrons confined inside a two-dimensional cavity, employing an STM. This surface quantum cavity or quantum corral (as it was called) has been built by using the same STM also as the manufacturing machine.
The experiment starts from a blank Cu(111) surface, held in atomically clean conditions in Ultra High Vacuum (UHV). The STM tip is used to position individual Fe atoms onto the Cu surface (the selection of Fe and Cu is by no means an evident one!). This is done by locating the tip above a "Fe atom storage place" somewhere at the surface, and applying a (positive) voltage to the tip: if done properly this makes one of the Fe atoms to become adhered to the tip. Next the tip is brought (with the atom on it) to the place where the



a. b.

*Figure 10.18. The direct imaging by STM of quantum interference effects and the formation of quantum eigenstates in a surface 2-dimensional electron gas at the Cu(111) surface, obtained in UHV conditions. The boundary conditions for the wavepatterns are defined by arranging Fe atoms on the surface employing the same STM. a. The positioning of a single Fe atom at the surface leads to the formation of a clear ringshaped wavepattern (13*13 nm image, I=1.0 nA). b. Arranging 48 Fe atoms in a ring forms a closed surface cavity or quantum corral; a beautiful standing wave pattern is set up, showing directly the spatial distribution of the probability of the eigenstates (ringdiameter 14.2 nm; I=1.0 nA).*

atom has to be positioned, followed by applying a negative voltage to the tip to "launch" the atom into place.
The peculiar feature of the Cu(111) surface is that a 2-dimensional electron system is formed at the surface. Evidently any additional atom at the surface will affect the eigenstates of the electron system, effectively via its change of the boundary conditions. This will become visible in the shapes of the wavefunctions, i.c. in the interference patterns of the electronic states. This is what the STM can measure.
Once an Fe atom is appropriately positioned the STM is switched to the measuring mode and the local DOS is scanned. Figure 10.18a shows the result. In addition to the large central peak showing the well-localised Fe atom a striking circular wave-pattern is set up in the surface electron system, demonstrating the interference of the injected electrons (from the STM tip) with the single Fe atom.
The next step in the experiment is to add more Fe atoms in order to increase the confinement of the electrons. More specifically, Eigler and Crommie (Science **262**, 218 (1993)) were able to arrange a total of 48 Fe atoms in a perfect ring, forming a so called quantum corral. Figure 10.18b shows the stunning STM picture. Inside the corral a very strong ring-shaped pattern is visible, thus forming one of the most direct experimental demonstrations of quantummechanics textbook examples. The actual pattern that is found evidently depends on the wavelength of the injected electron, and so on the energy or tip voltage; this is also found experimentally.

10.6 Closing remarks

The field of scanning probe methods and techniques is still evolving rapidly. New methods and applications become apparent at a high rate. No effort has been taken to describe this lively field in full, but we have hinted upon a number of interesting methods that are available and results that has been obtained over the last decade. One interesting recent development is the (scanning) near-field optical microscope ((S)NOM), which allows the detection or excitation of optical radiation (in the visible range) on a lateral scale of 50 nm and less (D.W. Pohl and L. Novotny, J. Vac. Sci Techn. **B12**, 1441 (1994)). As this also can be combined with STM the phenomena like electrofluerescence on a nanometer scale should be accessible. In addition the whole field of fabrication of structures and modification of surfaces on a molecular and atomic scale (as very briefly discussed in the last part of the preceding section, concerning the manipulation of Fe atoms on a surface) has not been discussed. Some more details on this and related subjects can be found in the literature given next to this section.

General references to the subject

1. "Particle Beam Microanalysis", E. Fuchs et.al., VCH Publishers,
New York(USA)/Weinheim(G)/Cambridge(UK), 1990
2. Populair iets over STM en AFM: Nature etc.
3. "Scanning Tunneling Microscopy I, II, III", Ed. R. Wiesendanger & H. Guentherodt,
Springer Verlag (Berlin), 1990/92/93
3. "Scanning Tunneling Microscopy and Spectroscopy", ed. D.A. Bonnel, VCH Publishers,
New York(USU)/Weinheim(G)/Cambridge(UK), 1993
4. P.K. Hansma & J. Tersoff, J. Appl. Phys. **61**, R1-R23 (1987)